DIVERSITY AND EVOLUTION OF THE IMMUNOGLOBULIN GENE SUPERFAMILY:

Thesis by

Tim Hunkapiller

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1993

(Submitted November 16, 1992)

ACKNOWLEDGMENTS

Whom not to acknowledge? This thesis, perhaps to an unusual degree, is the result of the many collaborations illustrated by the author lists of the following papers. Each of those individuals has been more or less significant to this effort. Of course, certain individuals do stand out either for their persistence, insight or moral support - most often for all three. Lee Hood must be noted for tolerantly allowing me to follow my own path for many years as well as providing an environment within which this random walk could be successful. This thanks includes all those who have shared tenure in Lee's lab over the years, providing the stimulating and sometimes combative atmosphere I enjoy. In a more general way, this acknowledgment is made to the entire Caltech community, particularly the exceptional members of the Division of Biology. Caltech is not only unique in the world in tolerating non-standard directions, it is just as unique as an amazingly stimulating intellectual environment. Particular note must be made, however, to individuals who have had the most direct impact on my career and whatever limited success I have had at Caltech. I would like to thank Mike Hunkapiller, Nelson Johnson and Mitch Kronenberg for introducing me to real science; Marie Malissen and Jane Parnes for respect and stimulating collaborations; Judy Barhydt and Mariel Clemenson for being good friends; Jim and Ellen Strauss for suffering through my growth as a software engineer. Special note is needed for my long-time collaborator and friend Joan Goverman. We've shared a lot of good science, good and bad times, and a close friendship. It is how science should be. This special mention is shared by Dana and David. Finally, I want to thank all of those people that ask me every time they see me "have you finished yet?" I hope so.

ABSTRACT

The Immunoglobulin Gene Superfamily is characterized by a common protein homology unit that is present in arguably the largest and most diverse set of genes and gene families of any protein motif. This distribution indicates that the homology unit is a remarkably versatile functional unit. Its central role in defining the complex phenotypes of the immune and nervous systems, likewise, is testament to the ability of the motif to support an amazing and unique degree of diversification. Understanding more about the function, structure and evolution of the Immunoglobulin Gene Superfamily can provide insights into both the general issues of complex system evolution as well as the specific nature of the various systems the superfamily plays a central role in. This thesis is a collection of work aimed at a more thorough understanding of these elements. Particularly, these works summarize much of our current understanding of the members of the Immunoglobulin Gene Superfamily along with speculations on their evolutionary history as well as both the evolutionary and somatic mechanisms responsible for their diversity. This work includes initial descriptions of several features relevant to somatic diversification of rearranging immune receptors, including: 1) the role of joining imprecision in the generation of junctional diversity in immunoglobulin kappa chain; 2) the initial description of the T-cell beta chain J/C locus; 3) the translation of T-cell beta chain D gene segments in all three reading frames; 4) the occurrence of a cryptic rearrangement signal in most rearranging V families; 5) the first description of the mechanisms of class switching between heavy chain mu and delta genes; 6) the limited diversity of germline T-cell beta chains; 7) the shared complementary determining region structure of T-cell beta chains and immunoglobulin heavy chains. Also, from these efforts, new members of the superfamily have been identified including MHC class I molecules, L3T4 and Myelin Associated Glycoprotein. Various observations concerning the evolutionary relationships of these molecules and motifs have been made. Particularly, a variation on the basic homology unit motif has been proposed that probably more nearly represents the primordial sequence and function.

As a result of these discoveries, a new, comprehensive picture of the immunoglobulin superfamily is emerging that has implications for interpreting current functional relationships in the context of the evolutionary history of the members.

Particularly, it is suggested from this work that the ability of the homology unit to accommodate diversity has made possible the evolution of the superfamily. Given the tremendous diversity within the superfamily, it might be assumed that selective pressures favoring diversity have driven its evolution. However, much of the analysis within this collection suggests that, on the contrary, diversity is an inherent feature of the conserved protein and gene structure of the homology unit and that it was the *a priori* diversity itself that drove and shaped the evolution of the complex systems that employ the homology unit today. This basic diversity is the consequence of three characteristics of the homology unit. First, the tertiary structure of the protein motif is such that homology units tend to interact preferentially to form homo- or heterodimers, forming the basis of many of the receptors and the receptor/ligand interactions common within the superfamily. These combinatorial associations increase both the somatic and evolutionary potential for diversification. This can lead to the rather sudden appearance of new functional associations between existing members of the superfamily preadapted for otherwise unrelated functions. Second, except for a minimal number of amino acid residues involved in critical intra- and interchain interactions, the primary structure of these units can vary dramatically and still provide for essentially the same tertiary structure. This has been borne out by various crystallographic studies. The variability is particularly true of the loop structures normally identified with antigen specificity, but seen in other extended families as well. Reduced constraints on structural sequences inherently promote the establishment of variation within populations. Third, with very few exceptions the genes of the superfamily, the homology units are not only encoded by discrete exons, but these exons have a shared 1/2 splicing rule. That is, each is begun with the second 2 bases of a codon and ended with the first base. This allows the in-frame splicing of any number of tandem homology units, while maintaining functional protein domains. This rule generally applies to the non-homology unit exons of member genes as well. This allows, through relatively simple genetic events, the development of new contexts for homology unit expression, both by simple expansion and contraction of homology unit number and exon shuffling. This is probably at work, as well, in the frequent occurrence and utilization of alternative transcripts seen throughout the superfamily. Many of the recognized occurrences of alternative splicing, such as that between membrane-bound and secreted forms, indicate that this gene structure provides for a further level of functional diversity and the expansion of the virtual genetic information.

Beyond the explicit discussion of the superfamily members, this work also speaks to various issues of evolution in general. In particular, the history of the superfamily suggests the importance of canalization and non-gradual episodes of evolutionary change. It can contribute, as well, to the discussion of adaptive versus neutral change.

TABLE OF CONTENTS

IV. INTRODUCTION

FOREWORD

Although members of the immunoglobulin gene superfamily have been identified mostly as the main effectors of the immune response, their evolutionary history appears to significantly predate the appearance of immune specificity. Figure 2 on page 10 illustrates the structural and functional protein sequence characteristic of this superfamily, the immunoglobulin homology unit. This motif was first recognized in the antigen receptor molecules of the humoral immune system, the antibodies or immunoglobulins, hence the name. However, the same basic motif is now known to occur in an array of molecules critical to the development of phenotypes as disparate from the immune receptors as the nervous system and cartilage formation. Homologous gene segments are found in both protostomes and deuterostomes, further indicating an ancient history. The primary function of these proteins as mediators of cellular interaction suggests a critical role in the evolution of multicellular organisms. Over fifty different gene families incorporating the superfamily motif have been identified, representing thousands of different gene segments. Figure 1 on page 12 represents the diversity of these families. The baroque nature and enormous potential for the somatic diversification of the immune receptors also provide a unique model for the evolutionary development of strategies for somatic adaptability.

This thesis is a collection of papers concerned with characterizing the genes and proteins of the immunoglobulin gene superfamily. It represents my efforts to more fully understand the evolution of this complex array of genes and associated phenotypes, particularly the vertebrate immune system. These efforts have included both theoretical and practical contributions. Also, along with the biology, considerable effort has been aimed towards developing tools that facilitate both the collection and understanding of the type of data needed to more thoroughly understand these as well as other molecular information.

This collection is organized into three sections. The first or introductory section includes, besides this foreword, two papers that specifically summarize the observations, conclusions and speculations to be found in the following papers. The first paper in this section is a particularly broad discussion of the immunoglobulin gene superfamily and, as well, provides a reasonably concise summation of the biological and evolutionary insight represented by the work in the following sections. The second paper was a review of the state of molecular T-cell biology at the time of its writing and provided a synthesis of a comprehensive model of T-cell antigen recognition based significantly on much of the work in the final section

Its most basic tenets concerning T-cell selection, activation and structural implications have held up well under further testing. These two papers were written explicitly as introductions to my work included in this collection and should be seen, hence, as the introduction and summation to this thesis.

The second section is a collection of papers that cover the more general issues of the immunoglobulin gene superfamily, providing an introduction to the structure/function implications of multigene families, the superfamily motif and its occurrence in diverse complex phenotypes. The papers in this section provide syntheses concerning the evolutionary implications of the superfamily as well as the identification of novel members and predictions regarding their function.

The final section is concerned with research papers directly describing the isolation and/or characterization of specific immune receptor genes and proteins. It is divided into four subsections. The first three contain contributions regarding the three main classes of immune receptors in T-cell antigen recognition and activation: T-cell receptors, molecules of the major histocompatibility complex (MHC) and the co-receptor molecules that work with T-cell receptors to define and aid in T-cell function (CD4 and CD8). The fourth subsection contains early experimental work that defined two novel forms of somatic diversification of immunoglobulins.

As indicated, the work presented here was done in collaboration with many other authors over an extended period. The general theme of all of these efforts has been a desire to understand the history and implications of the diversity of the Immunoglobulin Gene Superfamily. In this regard, my contributions to these efforts have established several new observations and conclusions regarding the mechanisms of diversity within the superfamily, including the following.

1) The combinatorial rearrangement of a set of V and J gene segments was likely to be the driving force of somatic variation of immunoglobulin chains. However, further diversification of immunoglobulin kappa chains likely resulted from variation introduced at the time of joining of the V and J gene segments during development. An imprecision of the joining site created variability from the germline of the codons included in the final protein coding transcript. This observation relied on protein sequencing of related kappa chains and was subsequently verified with the characterization of the germline sequences. This mechanism has been found to be a significant diversifying factor in all other known rearranging gene segments of the immune system (Sec. VI.4.1).

2) The first analysis of the T-cell beta chain constant region locus found 2 C regions, each with a cluster of 6 J gene segments. V rearrangements were possible to both clusters of J segments, using the same basic rearrangement signals as immunoglobulins (Sec. VI.1.1).

3) Diversity of the T-cell beta chain is further enhanced by the ability to join the D gene segment in all three reading frames. This was the first observation of a sequence coding for protein in all reading frames. A similar, but more limited use of the same mechanism was subsequently found in immunoglobulin heavy chains (Sec. VI.1.2).

4) A region of protein coding sequence conserved in most rearranging V gene families was likely able to serve as a secondary rearrangement signal for V/D and V/J joining. The conservation was due to both protein sequence constraints as well as codon preferences resulting from eukaryotic distaste for CG dinucleotides. Subsequently, rearrangements were discovered that used these sequences and illustrated their role in rescuing nonfunctional primary rearrangements. This observation is the foundation for a model on the evolution of the D gene segment from existing V genes (Sec. V.5).

5) During differentiation of B cells, the same Variable regions become associated with different Constant regions in a process called class switching. For the expression of mu and delta constant regions, we were able to show that both alternative splicing of long transcripts that included both constant regions as well as actual rearrangement events that repositioned the V to the Delta constant region were involved. This provided for the association of the same antigen recognition element with different effector functions (Sec. VI.4.2).

6) There was significant question as to how T-cell receptors were able to recognize both antigen and molecules of the MHC. Complicated models of dual recognition sites were proposed. My analysis of the distribution of the variability of T-cell V beta chains, however, indicated that they employed the same complementary determining regions that immunoglobulins did. This, along with a detailed analysis of the conserved residues of both receptor families lead to the currently accepted view that both receptors share the same basic structure and antigen recognition mechanism (Sec. VI.1.2).

7) Diversity of the T-cell receptor was assumed to be greater than that of immunoglobulins. We were the first to show that, if this were true, it was not a result of more germline segments as T-cell receptors with non-overlapping specificities employed the same V beta chain. Most significant, these analyses rebutted the widely held view that antigen and MHC specificities must be relegated to either the alpha or beta chain exclusively (Sec. VI.1.3).

Along with variation and diversification of rearranging immune gene segments, several more general observations have been made with regard to other members of the superfamily, including the following.

1) A randomly isolated brain transcript was described by another lab. My analysis discovered it to be a member of the superfamily. It was subsequently found to be Myelin Associated Glycoprotein (MAG), a major cell adhesion molecule and perhaps involved in cellular homing (Sec. V.3).

2) Analysis of Neural Cell Adhesion Molecule (NCAM) indicated that the several immunoglobulin homology units of the initially incompletely characterized transcript were, in fact, likely to be the recognition elements responsible for homodimeric interactions. Further characterization of the gene confirmed this (Sec. V.3).

3) Analysis of NCAM, MAG and several other non rearranging members of the superfamily resulted in the description of a third major family of homology units (H) along with the V and C families. This was simultaneous with another author (Sec. IV.2).

4) Class I and Class II MHC sequences present antigen to T-cell receptors. Class II receptors are heterodimers of alpha and beta chains while Class I receptors have a heavy chain associated noncovalently with a small light chain , beta-2 microglobulin. Novel computational tools were applied to the analysis of the Class I sequences to discover and define statistically the likelihood that they were members of the superfamily. The first analysis of the mouse Class II E-alpha and A-beta sequences along with the genomic structure of the genes suggest that alpha chains and beta-2 microglobulin are related, while the beta chains appear more similar to the Class I heavy chain (Sec. VI.3.1-3).

5) L3T4 (CD4) is a coreceptor for T-cell recognition and activation involved in the recognition of MHC Class II. Analysis discovered it be constructed of multiple homology units analogous to NCAM and MAG, but with interesting truncations of the motifs. Analysis of T8 (CD8), the analogous coreceptor for MHC Class I sequences indicated a single motif structure that was significantly related to V gene sequences and not L3T4, indicating independent derivations (Sec. VI.2.1-2).

6) MRC OX-2 is a cellular marker of unknown function, but with a distribution similar to another superfamily member, THY-1. Analysis indicates not only a V-like homology unit, but also a relict J-like region even though OX-2 does not rearrange. This, along with the subsequent discovery of a similar J-like sequence in Lyt2 beta (mouse CD8), has interesting implications for the evolution of the multisegmented immune receptors (Sec. V.5).

ADVANCES IN IMMUNOLOGY, VOL. 44

Diversity of the Immunoglobulin Gene Superfamily

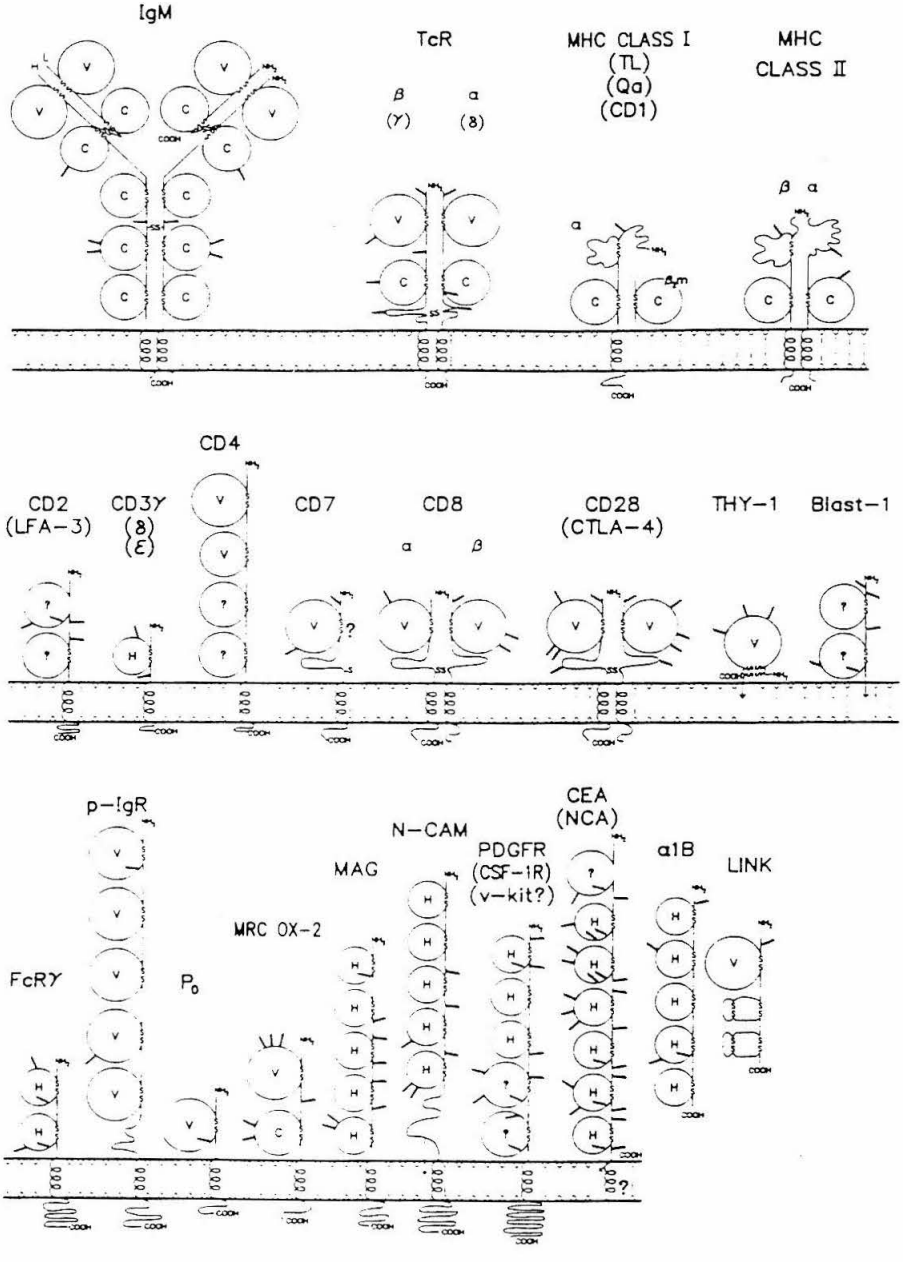
TIM HUNKAPILLER AND LEROY HOOD

Division of Biology
California Institute of Technology,
Pasadena, California 91125

I. Introduction

The vertebrate immune response consists of a complex set of cellular and serological reactions that provide protection from foreign and abnormal self-macromolecular structures (antigens). These responses are mediated by an equally complex array of immune recognition elements. During the past decade, many of these molecules have been characterized and have been found to share a common evolutionary precursor, the immunoglobulin homology unit (Cunningham *et al.*, 1973; Strominger *et al.*, 1980; Williams, 1984; Hood *et al.*, 1985). Recently, several molecules with no known immunological functions have also been shown to share this same precursor element (Hunkapiller and Hood, 1986; Williams, 1987). Together, the genes encoding these related molecules have been defined as the immunoglobulin gene superfamily (IgGSF) and include both multigene and single-gene representatives. These IgGSF products represent an amazingly diverse array of functions from immune receptors to cartilage formation, reflecting the versatility of the shared common structure (see Fig. 1).

We summarize here our understanding of the members of the IgGSF and speculate particularly on their evolutionary history as well as both the evolutionary and somatic mechanisms responsible for their diversity. We define here a new variation on the homology unit motif that probably more nearly represents the primordial sequence organization than either of the two previously defined units. We believe that the unique ability of the immunoglobulin homology unit to accommodate diversity (information) has made possible the evolution of the complex phenotypic traits of the IgGSF. We also suggest that the diversity potential of the informational multigene families (Hood *et al.*, 1975) of the immune receptors is generally more important evolutionarily than the individual strategies used to generate it. Thus, many different, but selectively equivalent, diversifying strategies are seen in the immune receptor gene families of the IgGSF. Finally, the genetic organization of IgGSF


IgM
TcR
β
(γ)
α
(δ)
MHC CLASS I
(TL)
(Qa)
(CD1)
MHC
CLASS II
β α
α
β α
CD2
(LFA-3)
CD3γ
(δ)
(ε)
CD4
CD7
CD8
α β
CD28
(CTLA-4)
THY-1
Blast-1
p-IgR
CEA
(NCA)
N-CAM
PDGFR
(CSF-1R)
(v-kit?)
α1B
LINK
MAG
MRC OX-2
FcRγ
P0

members and gene families as well as structural and interactive properties of their proteins impart unique possibilities for their further evolutionary diversification. Recently, a review on the IgGSF has appeared elsewhere (Williams, 1987).

II. The Immunoglobulin Homology Unit and the IgGSF

We define a superfamily as a series of genes that share an evolutionary homology (i.e., common ancestor), but do not necessarily share function, genetic linkage, or coordinate regulation (Hood *et al.*, 1985). Members of the IgGSF have been defined by the presence of one or more regions homologous to the basic structural unit of immunoglobin (Ig) molecules, the Ig homology unit (Hill *et al.*, 1966). These units are characterized by a primary sequence about 70–110 residues in length with an essentially invariant disulfide bridge spanning 50–70 residues and several other relatively conserved residues involved in establishing a tertiary structure referred to as an antibody fold. Two basic homology unit types have been defined from crystallographic analysis of the variable (V) and constant (C) regions of Ig (see below). The tertiary structure of a V region is dominated by a series of nine antiparallel β strands, connected by variable-length loop sequences, that assume a characteristic barrel or sandwichlike structure with two β sheets, stabilized by the disulfide bridge (Fig. 2) (see Amzel and Poljak, 1979). There are four β strands in one sheet and three in the other. The extra pair of β strands is essentially situated between the faces of the sandwich. The β strands are characterized by alternating hydrophobic and hydrophylic amino acid residues. The hydrophobic side chains are oriented toward the interior and help stabilize the interaction between the two sheets. The outpointing hydrophylic residues mediate the interchain interactions. The disulfide bridge further stabilizes this basic structure, providing compact, globular domains that are relatively proteolytically insensitive.

FIG. 1. Schematic diagrams for the members of the IgGSF. Disulfide bonds are represented by (s–s). Homology units are indicated as loops labeled V, C, or H (see text). Loops of uncertain relationship to homology units are labeled with a question mark. Different-sized loops illustrate the relative differences in length between the conserved disulfide bond of the labeled types. Membrane-spanning peptides are shown as simple helices. Glycophospholipid linkage to the membrane is represented as an arrow. Intracytoplasmic regions are drawn with wavy lines that indicate their relative lengths. Extra- and intracellular orientations are indicated by NH_2 and COOH labels on the protein chains, respectively. Possible asparagine-linked carbohydrates are shown as jagged lines extending from the protein chains. Note that these sites are not necessarily conserved between alleles or across species, but are representative of at least one known example of the labeled protein. Related IgGSF members are illustrated with a single structure, as indicated by the name labels above each structure.

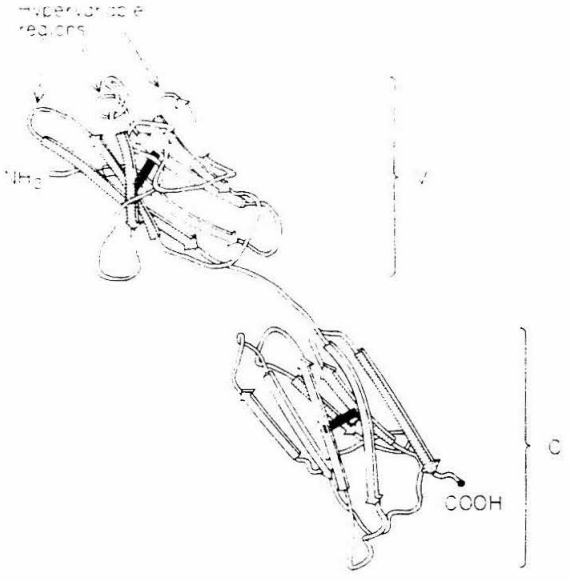


FIG. 2. Tertiary organization of V and C homology units (after Edmundson *et al.*, 1973). This illustrates the tertiary structure of an Ig light chain. The β strands and their orientation are shown as flat arrows. Opposite faces of the sandwich structure are indicated by either blank or hatched β strands. The disulfide bond is indicated with the solid bar.

The C region units lack the pair of internal β strands (Fig. 2), but otherwise assume the same general structure with a distinct, but overlapping, series of conserved residues. The lack of this extra pair of β strands decreases the distance between the two cysteine residues of C regions relative to those of V regions. The extra loop sequence connecting these two strands in V regions is critical to the formation of the antigen-binding pocket of antibodies (see below). Though there are enormous variations of primary sequences used, the basic scaffold structure of the antibody fold appears to be highly conserved within the IgGSF members. As new members of the IgGSF have been characterized, their homology units have generally been defined as either V- or C-like based on primary sequence similarities and secondary structure predictions. Many of the more recently discovered members have a primary and secondary structural motif that, although shared among themselves, does not preferentially fit either the previously defined V or C homology units. It is a more compact unit, even shorter in length between the two cysteines than most C-like units. Accordingly, we define a new class of homology unit, denoted H, to encompass these members. This third type of

homology unit has also recently been defined as C2 (constant 2) mostly due to its lack of the V-defining extra β-strand pair (Williams, 1987). Overall, however, this unit appears to have a relatively equidistant relationship to both V and C homology units. Therefore, we prefer a more generic definition that does not suggest any functional relationship to one or the other homology unit types and emphasizes the building block nature of this unit. Because both V and C units are each more closely similar to H units than to each other, the H unit probably reflects a more primordial motif (see below), suggesting that the original members of the superfamily arose in early metazoa and clearly carried out cell-surface recognition functions unrelated to vertebrate immunity. Figure 3 represents the consensus structures of each of the three homology unit types.

To discuss this or any other superfamily, criteria must be defined that establish the likelihood that any particular sequence is a homolog to the superfamily. When attempting to define evolutionary homology between any two sequences, various mathematical calculations can be made to advance arguments of statistically significant similarity (e.g., Dayhoff *et al.*, 1983). These methods rely on defining the similarity or distance score between any two protein sequences, calculated as a

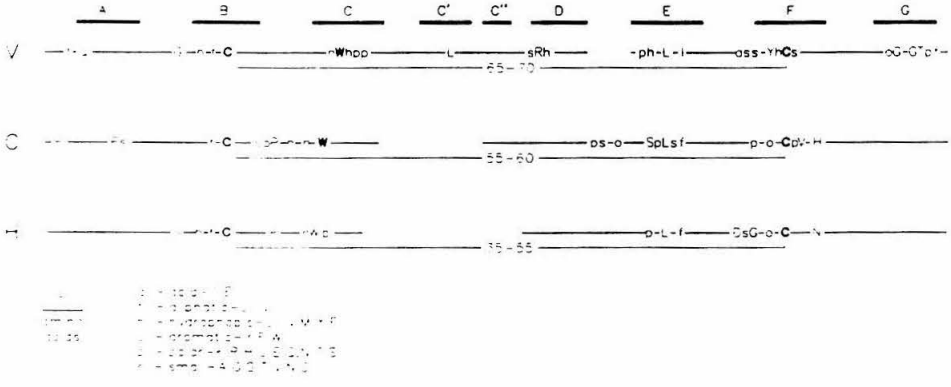


FIG. 3. Sequence motifs of each homology unit type. This figure illustrates idealized aligned examples of each homology unit type. Representative amino acid residues of each type are shown in their relative position along the polypeptides. Uppercase letters are single-letter amino acid code. Lowercase letters represent a selection of amino acids with similar functional or physical properties. The key to this code is shown in the lower left. Relative gaps are introduced into each chain to align residues conserved between types. The more nearly invariant cysteine and tryptophan residues are highlighted. The distances between the conserved cysteine residues are shown beneath each chain. The approximate locations of the β strands are illustrated with heavy bars above the sequence representations. Note the two β strands (C′ and C″) not found in the C and H domains. It is also clear that β strand D may not be found in the shorter H domain sequences.

function of the degree of sequence identity between them. A significant relationship is usually defined by a similarity score a minimum number (e.g., >2-3) of standard deviations (SD) above what would be expected from comparison of two random sequences. The expected value is generally calculated by randomizing or "shuffling" one or both sequences many times (100-150) and each time recalculating their similarity. It is the mean of these scores that forms the basis for the reference score for "unrelatedness." The greater the deviation is from this standard, the closer the relationship. Therefore, these values should reflect the evolutionary history and particular relationships within a family of related sequence, i.e., genealogy. These methods work well for relatively recently diverged sequences whose differences can be statistically viewed within a model of gradual sequence divergence at relatively constant rates. However, selective constraints against further sequence divergence between more distantly related sequences disrupt this model. For example, the conserved structure of IgGSF members is apparently mediated by a limited, but selectively conserved, set of amino acid residues at particular positions in the primary sequence, while many of the rest of the sequences are allowed to vary dramatically. There must be, then, a limit to the apparent sequence divergence between any two sequences such that they can maintain shared structural features. That the number of explicitly conserved amino acid residues need not be large is illustrated by the fact that even proteins generated by genes belonging to the same family may be less than 20% similar. However, many of the remaining residues are relatively conservative substitutions for the same position in another member. Consequently, similarity scores between proteins with variably selected residues decrease linearly with actual evolutionary distance only within a relatively restricted distance range. In other words, as evolutionary distance increases, distance scores tend to level out within a relatively narrow (and low) range, thus losing any fine distance information. Within this range ($\sim$2-6 SD) significant noise is generated by localized regions of presumably convergently similar sequences. This is because no protein sequence can be viewed as legitimately random; i.e., not all possible combinations of linear and tertiary relationships between amino acids are biologically meaningful. Hence, linear sequences that successfully define a limited number of structural motifs must convergently (independently) arise much more frequently than statistical analyses would predict from amino acid composition alone. Thus, relatedness scores between sequences that share analogous structural features will tend to overestimate their actual evolutionary relatedness. In the case of the IgGSF this is particularly true because of the dominant structural role of the β strands. The alternation of hydrophobic and hydrophylic residues certainly imposes a bias in these sequences for

the presence of certain amino acid combinations. Presumably any sequence that relies heavily on β-sheet structures will, by selective constraint, be biased toward a similar amino acid composition. Therefore, the contribution of covergent evolution cannot be ignored when considering the relationship of distantly similar sequences. Another complication that exacerbates the difficulty in distinguishing divergent from convergent relationships is the relative overrepresentation of IgGSF sequences in the data bases that new sequences are generally compared against. In the absence of truly significant matches, a data base comparison of almost any sequence will generate a list of "best" matches that includes at least some members of the IgGSF because there are so many diverse IgGSF examples in the data bases.

To identify significantly related sequences it is critical to start with reasonable sequence alignments. Alignments are the means by which residue-to-residue or base-to-base homologies are established. This means that the various penalty and weighting factors used to calculate the distances implied by the substitutions, deletions, and insertions of a particular alignment must reflect biologically and structurally "appropriate" assumptions about the likelihood of the various genetic events as well as the selective pressures for maintaining structural integrity. In this regard, it is important to take into consideration established sequence and structural homologies when evaluting the addition of a new member to the IgGSF. It is meaningless, for example, to align a new sequence to two established IgGSF members in ways that maximize its similarity to each of the two independently, but would result in a misalignment of already well-defined relationships between the two established members. Therefore, it is important that all family members fit a consensus alignment. In the IgGSF this consensus alignment must take into consideration the β-strand/loop organization and the repertoire of conserved amino acid residues. As reflected in the voluminous collection of V- and C-like sequences aligned by Kabat and Wu (1987), not only substitutions but length heterogeneities (deletions/insertions) are overwhelmingly concentrated in the loop regions of IgGSF sequences, presumably because they have less selective impact on the basic structure. Therefore, while indiscriminate use of gaps during alignment of any two IgGSF members (or potential members) may optimize the statistical score between them, it may ignore the selective constraints dictated by biological considerations.

Unfortunately, the evolutionary distance between the major groups of the IgGSF and the nonrandom modes of modification of its members make standard methods of statistical comparison singularly unreliable for establishing not only the broader relationships within the superfamily but also for determining whether a sequence belongs to the IgGSF at

all. For example, one attempt to calculate the relative distances between various distantly related IgGSF members using standard techniques (Williams *et al.*, 1987) determined that the Thy-1 molecule (a T cell surface antigen; see below) had a similarity score 4.2 SD above random with the fibrinogen γ chain and a score of 6.6 SD with the poly(Ig) receptor (see below). These scores represent highly significant p values of about 10^{-5} and 10^{-9}, respectively. However, the fibrinogen γ chain and the poly(Ig) receptor had a score of only 0.8 SD, exhibiting no significant relatedness. Except in unusual circumstances, a syllogism between scores should be maintained, i.e., if sequence A and sequence B are related to sequence C, then A and B should also be related. If these relationships are lacking, then one must question whether, in the example given above, any significant evolutionary relationships are reflected. Accordingly, other criteria must also be considered when attempting to establish IgGSF relationship(s). For example, secondary and tertiary protein structure conservation can be critical when considering the relationships within a family of molecules so strongly dependent on a conserved structure. However, there is no three-dimensional structural information for most IgGSF members, only calculated predictions based on less-than-perfect methods (Novotný *et al.*, 1984). Also, there are the caveats concerning the possibility that presumably unrelated sequences can convergently assume a similar tertiary structure (e.g., Richardson *et al.*, 1976). Therefore, we have taken a broadly cladistic approach to defining members of the IgGSF and their relationships, based upon a series of hopefully reasonable, objective, and subjective predefined taxa of features and assumptions, no single one of which is absolute. Primarily, a sequence must have a consistent, statistically significant degree of primary sequence identity to many, divergent representatives of the IgGSF. In addition, a sequence must also minimally maintain most of a small series of particularly conserved residues (Fig. 3) and a predominant β-strand potential throughout its sequence for us to consider it a member of the IgGSF. Also, direct or functional association with other, better defined homology units can lend subjective support to considering a more ambiguous sequence as a relic unit, even if it does not meet enough of the more objective criteria alone. Several of the larger, polydomain sequences (to be discussed later) exhibit sequence regions that can be argued to be derived evolutionarily from IgGSF sequences, given this latter assumption. Another striking general feature of IgGSF members, also useful in defining IgGSF membership, is that individual homology units of IgGSF members are encoded by discrete exons (with a few exceptions to be discussed later). Moreover, splicing between these exons always occurs between the first and second base of the boundary codon (½ rule). Other factors such as percentage identity, the ability to undergo somatic DNA rearrangements, the particular gene segments

used in the rearrangement process, the comparative exon organization of sequences, the functional roles and associations, and the chromosomal linkages can be considered to refine further the genealogical relationships. The sequences discussed in this paper all comfortably conform to this set of rules and are all generally accepted as members of the IgGSF.

III. Receptors of Immune Recognition

Any discussion of the diversity of the IgGSF must, by weight of data, be dominated by the molecules that mediate specific recognition of antigens: the Ig of B cells, the antigen-specific receptors of T cells (TcR), and the class I and class II proteins of the major histocompatibility complex (MHC). Ig can be expressed either as cell surface-bound receptors or as secreted humoral antibodies. Ig are capable of recognizing soluble or "free" antigens and play a role in both the cell surface triggering of B cell differentiation and as the effector molecules of humoral immunity. TcR, in contrast, occur only as surface-bound molecules and are able to recognize antigen only when it is presented in association with a class I or class II MHC molecule on a target cell. Hence, T cell recognition is said to be MHC restricted (see Schwartz, 1984). The T cells generally mature in the thymus and it is there that self-reactive T cells are eliminated and the processes that confer MHC restriction occur (Bevan, 1981). In general, class I molecules present antigen to cytotoxic T cells while class II molecules present antigen to regulatory or helper T cells. TcR do not appear to mediate directly the effector responses that occur upon T cell activation. Rather, they are present in a matrix of accessory molecules (CD3, CD4, CD8, etc.) that are involved in T cell activation and responses (see Townsend, 1985).

The basic Ig structure is a tetramer constructed of two identical, disulfide-linked heterodimers, each composed of one light (L) and one heavy (H) chain. Both chains are divided into an N-terminal V and a C-terminal C region. This interaction is homophylic in the case of the C-terminal C_H domains and heterophylic between the V regions and the C_L and C_{H1} regions. The V and C_L regions are single homology units while C_H regions consist of two to four related, but distinct, units tandemly linked. Pairs of similar homology units between chains associate to generate globular domains which represent the functional units of the molecules. The V and C homology units use opposite faces of the sandwich structure to interact with like units through mostly β-sheet interactions. Interchain association is often stabilized through disulfide bonds. Together, the V regions from both chains form the antigen-binding domain while the C region domains mediate various effector responses. As will be discussed later, many of the C-related functions

are initiated through further heterophylic associations of the C domains with homology units of other IgGSF members. Ig are humoral or membrane-bound depending on which exons of the C_H genes, i.e., those responsible for secretion or those responsible for membrane insertion, are included in the mature RNA by alternate pathways of RNA splicing (Early *et al.*, 1980a). The antigen-binding site of an Ig V domain is derived mostly from three or four relatively "hypervariable" (HV) loop sequences from each chain that connect the β strands of the more conserved "framework" regions (Wu and Kabat, 1970; Capra and Kehoe, 1974). The second HV region is generated by the connecting loop of the V-specific pair of β strands (Amzel and Poljak, 1979).

The predominant TcR are heterodimers, composed of an α and a β chain (see Hannum *et al.*, 1984; Meuer *et al.*, 1984). The α/β TcR have been found on all functional cytotoxic T cells and helper T cells. Together, they appear to be necessary and sufficient to confer specificity for antigen and MHC (Yagüe *et al.*, 1985; Ohashi *et al.*, 1985; Dembić *et al.*, 1986; Kuo and Hood, 1987). Each receptor is associated in the membrane with CD3, a multimeric protein complex of four or five polypeptides that appears to have a transducer role in T cell activation (see Terhorst *et al.*, 1986). Molecular analyses have shown that both the α and β chains have a single V and C homology unit, a connecting sequence of unknown function, a transmembrane, and a cytoplasmic region (Fig. 1) (see Davis, 1985; Kronenberg *et al.*, 1986). The C_α homology unit has a highly unusual structure apparently lacking an internal β strand and a highly conserved tryptophan residue important in stabilizing the overall fold. It is assumed to be evolutionarily derivative of the C unit lineage rather than that of H units, because of a gene and protein organization analogous to other immune receptor C units (see below). Although the sequence similarity between T and B cell V regions is low, the V regions of both receptors share about 15 conserved residues (Kronenberg *et al.*, 1986) that have been shown in Ig to be critical sites for determining three-dimensional structure through intra- and interchain interactions (see Amzel and Poljak, 1979). Various theoretical calculations of structure potential also suggest that the V domains of both receptors share essentially the same three-dimensional form (Goverman *et al.*, 1986; Novotný *et al.*, 1986). However, a reduced intracysteine distance in both V_β and V_α relative to Ig sequences may indicate a generally shallower (flatter) binding site (Goverman *et al.*, 1986). The total variability between the known V_β sequences examined so far is greater than that of the available Ig chains (Patten *et al.*, 1984), and this variability is more distributed throughout the sequence. However, analysis of extensive sequence data indicates that there are HV regions in V_β sequences at

positions homologous to those of Ig chains (Barth *et al.*, 1985; Behlke *et al.*, 1985; Concannon *et al.*, 1986). High background variability leads to ambiguous HV results for the V_α sequences (Arden *et al.*, 1985; Becker *et al.*, 1985; Hayday *et al.*, 1985a). The broader distribution of variability, combined with possibly flatter loop structures, may indicate that a generally larger area of the TcR is available to interact with antigen than is available in many Ig. Although Ig can also have relatively flat interactive surfaces with antigen (Amit *et al.*, 1986; Colman *et al.*, 1987; Sheriff *et al.*, 1987), such a general trend in TcR may have implications for the nature of the antigenic surfaces they recognize. Ig interact primarily with epitopes (antigenic sites) that are defined by a tertiary relationship between amino acids of a protein rather than their linear relationship along the sequence. It seems that the most flexible local structures of proteins may be generally more antigenic than the stiffer ones, as they would be able to adjust their conformation and generate a structure that better fits an Ig binding pocket (Westhof *et al.*, 1984; Fieser *et al.*, 1987). It has been suggested that an analogous flexibility of the Ig binding site could reciprocally facilitate such interactions (Colman *et al.*, 1987). The extremes seen in Ig, from the deep binding pockets of antihapten Ig to the flat surfaces of antiprotein Ig, might demonstrate a steric accommodation of the loop structures to the antigen with which they are bound more than some general binding-pocket geometry. TcR are not generally thought to interact with tertiary epitopes, but rather short linear sequences presented in a planar motif (see below). Hence, there would be less need for TcR as opposed to Ig to adjust to the variable shape of the antigen. However, it seems likely that there are no fundamental differences in the antigen-binding structures of Ig and TcR. This implies that they also share basic recognition strategies (Goverman *et al.*, 1986; Novotný *et al.*, 1986).

A second class of TcR also associated with the CD3 complex has a heterodimeric structure consisting of γ and δ chains (Brenner *et al.*, 1986; Bank *et al.*, 1986; Ioannidis *et al.*, 1987). This class of TcR is found on early thymocytes and about 1-5% of peripheral T cells (Lew *et al.*, 1986; Borst *et al.*, 1987; Moingeon *et al.*, 1987; Pardoll *et al.*, 1987; Bluestone *et al.*, 1987). The γ chain is more similar in overall sequence and organization to the TcR β chain (Saito *et al.*, 1984; Kranz *et al.*, 1985; Lefranc and Rabbits, 1985; Murre *et al.*, 1985). The δ chain is much more similar to the α chain, including sharing the deletion of the same internal β strand (Chien *et al.*, 1987a). The role of a γ/δ TcR is unknown, but, as mentioned above, neither the γ nor the δ chain is necessary to confer antigen or MHC specificity. The concordance between the structures of γ and β chains and δ and α chains does,

however, seem to imply a complementary function. The γ/δ TcR might define an important functional subset of T cells, with either or both a unique target repertoire or functional response. Some γ/δ-expressing cells can apparently exhibit both nonspecific and MHC-linked killing (Matis *et al.*, 1987).

A. Organization, Rearrangement, and Diversification of Ig and TcR Genes

Ig chains are encoded by three unlinked gene families, the heavy chain gene family and the κ and λ light chain gene families (Fig. 4) (see Honjo, 1983). In each family, V gene formation occurs through a process of gene segment rearrangement. The V_L genes are constructed by the juxtaposition of a V gene segment and a joining (J) gene segment, while a V_H gene is generated from the joining of a V, a diversity (D), and a J gene segment. RNA processing removes the intervening sequence between the rearranged J gene segment and the C gene. TcR α and β chains are encoded similarly, with V_α genes employing V and J gene segments and V_β genes constructed from V, D, and J gene segments (Fig. 4) (see Davis, 1985; Kronenberg *et al.*, 1986). The organization and sequence of the γ gene are most similar to those of the λ light chain genes, and do not appear to employ D segments (Fig. 4) (Hayday *et al.*, 1985b). The δ gene, while more similar to the α gene, does include at least two D segments (Fig. 4) (Chien *et al.*, 1987b). More interestingly, the δ gene family is entirely located within the α locus. The C_δ gene is located 5′ of the J_α gene segments and appears to rearrange both with uniquely V_δ gene segments and with V_α gene segments.

DNA rearrangements of Ig and TcR gene segments are mediated by specific recognition sequences proximal to the gene segments in complementary orientation (Fig. 5) (see Honjo, 1983; Kronenberg *et al.*, 1986). These sequences consist of a conserved heptamer linked by 12 or 23 nonconserved nucleotides to an A/T-rich nonomer. A recognition sequence with a 12-base spacer joins to one with a 23-base spacer and vice versa (Fig. 5) (Early *et al.*, 1980a; Sakano *et al.*, 1980). Transfected TcR D_β and J_β gene segments can rearrange appropriately in B cells (Yancopoulos *et al.*, 1986), and Ig/TcR hybrid rearrangements are known (Baer *et al.*, 1986). Thus, the same or similar enzymes are used for B and T cell gene rearrangement. Tissue-specific rearrangements probably reflect the tissue-specific accessibility of the local chromatin structure of the loci.

B. Mechanisms of Somatic Diversification

The diversity of both Ig and TcR must complement the range of potential antigenic determinants. An amazing array of mechanisms is employed

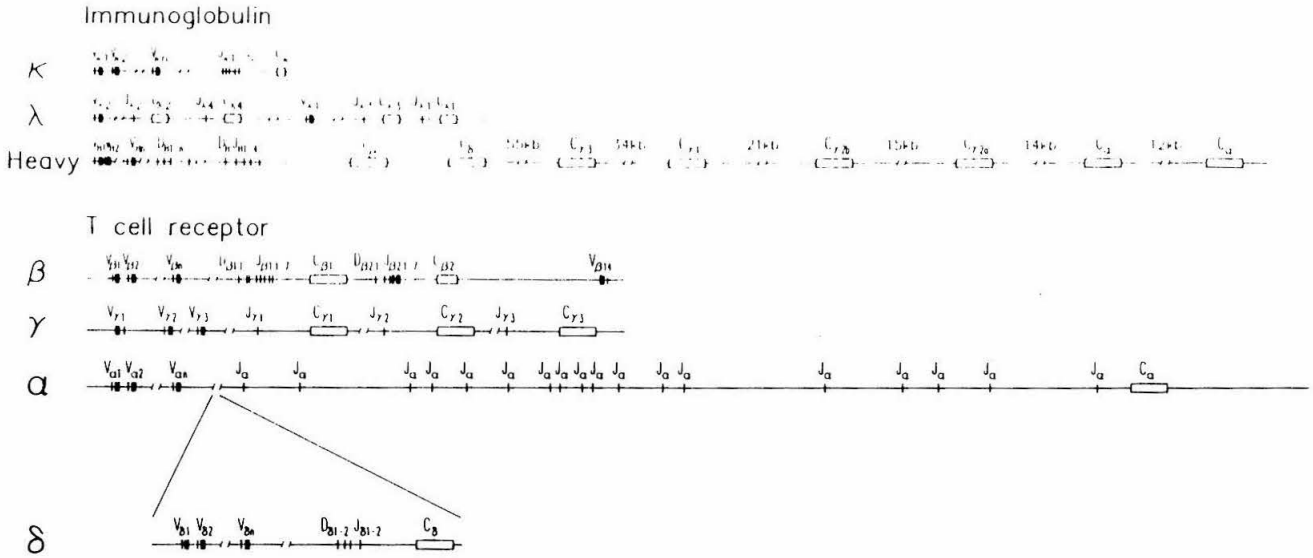


FIG. 4. Organization of the mouse Ig and TcR gene families. Open boxes are C-region genes and filled boxes are V-region gene segments. Vertical lines illustrate D and J gene segments. Numbers between C-region genes in the heavy chain family are distances in kilobases. Double slashes without numbers indicate unknown distances. Human gene families are organized similarly. The δ gene family is shown as an insert to the α gene family.

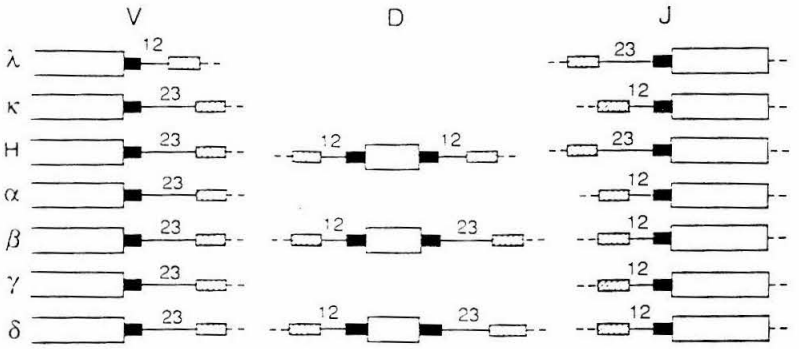


FIG. 5. Organization of V gene rearrangement signal sequences. The rearrangement signals for V, D, and J gene segments are shown in the relative orientation in which they are found in each gene family. The conserved heptamer is shown as a filled box and the nonomer is an open box. The distance in bases between each heptamer and nonomer is indicated.

to provide it. The basis of mammalian receptor diversity is the germline repertoire of gene segments and the consequent combinatorial possibilities that arise at both the DNA and protein levels during differentiation of both B and T cells (Table I) (see Honjo, 1983; Kronenberg *et al.*, 1986). At the DNA level there is a combinatorial rearrangement of gene segments within a family (e.g., any V_x can join to any J_x). At the protein level a combinatorial association may occur between the heterodimers (e.g., any α may associate with any β). Particular immune receptor loci may also have additional "nonstandard" combinatorial possibilities. The D_β and D_δ gene segments are flanked by asymmetric rearrangement signals, suggesting the possibility of D-D and V-J joining according to the 12-23 joining rule (Fig. 5). Possible examples of both events have been described for the β chain (Yoshikai *et al.*, 1984; Concannon *et al.*, 1986), but their frequency is unknown. However, the occurrence of D_δ-D_δ joinings is well established (Chien *et al.*, 1987b). The V_H gene segments have been shown to rearrange into already complete V_H genes at an internal sequence near the 3′ end of the V_H gene segment that mimics the 5′ rearrangement heptamer of D gene segments (Kleinfield *et al.*, 1986). This results in a combinatorial potential between any V_H gene segment and the very 3′ sequence of any other downstream V_H gene segment. Interestingly, all four families of TcR V gene segments conserve the same cryptic rearrangement signal sequence (data not shown) and, accordingly, have the potential for similar secondary rearrangements.

Two other diversifying mechanisms directly linked to the rearrangement process operate to extend the basic combinatorial diversity of both

TABLE I

DIVERSITY POTENTIAL OF IMMUNE RECEPTOR GENES

Gene	Heavy	κ	α	β	γ	δ
V gene segments	500	200	100	30	7	15
(subfamilies)	(8)	(5)	(16)	(17)	(5)	(8)
D gene segments	15			2		2
J gene segments	4	4	>50	12	2	2
D segments in three reading frames	+	Na	Na	+ +	Na	+
N region sequence	+ +	−	+	+ +	+ +	+ +
Junctional diversity	+ + +	+	+ + +	+ + +	+ +	+ + +
Somatic hypermutation	+	+	−	−	−	?
Nonstandard rearrangements	+	−	−	+	−	+ +
Combinatorial joining	V × D × J	V × J	V × J 100 × 50	V × D × J (30 × 3 × 12) + (30 × 3 × 6)	V × J 7 × 2	V × D × J (15 × 2 × 2) + (15 × 2)
Total	3×10^4	800	5000	1620	14	90
	×			×		×
Combinatorial association	2.4×10^7			8.1×10^6		1.3×10^3

B and T cell sequences. First, the actual site of joining can vary across a range of several coding nucleotides proximal to the heptamer in both segments (Fig. 6) (see Tonegawa, 1983). This results in the generation of length heterogeneities and, hence, significant diversification of the protein products. However, this flexibility also results in many V and J gene segments joined out of reading frame with respect to each other and, hence, "nonproductive" V genes. Each rearrangement appears to have only the randomly expected one-in-three chance of joining the segments in frame (Alt *et al.*, 1984; Lewis *et al.*, 1985). This possibility suggests that joining efficiency in both B and T cells has been sacrificed for the diversity contributed by a flexible joining process. It has been suggested that secondary V rearrangements into the cryptic rearrangement signal may "rescue" some fraction of these nonproductive V genes (Kleinfield *et al.*, 1986).

A consequence of joining flexibility is that the D_H and D_β gene segments can be used in all three reading frames, essentially multiplying the size of each D segment family by three (Goverman *et al.*, 1985; Kaartinen and Makela, 1985). However, D_H gene segments appear to be more restricted in the reading frames they employ than are D_β gene segments. This distinction cannot be accounted for by termination codons in the unused frames. It is unclear, then, whether this distinction is a consequence of specific differences in the rearrangement process, of selection at the protein level, or simply of a sampling bias (most characterized TcR sequences have been isolated without regard to specificity, unlike

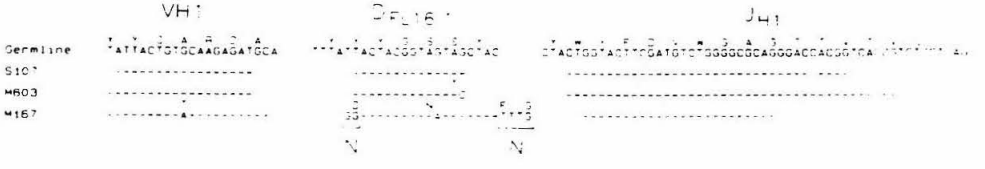


FIG. 6. Somatic diversification of Ig gene elements. Shown are the 3′ sequences of three related heavy chain V genes that each use the same V, D, and J gene segments (see Kim *et al.*, 1981, for sequence references). This figure is a simple illustration of how somatic diversification mechanisms can generate different proteins even if the same germ line information is used. Each rearranged V gene is listed beneath the representative germ line elements from which they were constructed. Dots indicate base-to-base similarity of each V gene to the germ line sequences. Translation of the germ line sequences are shown directly above in single-letter amino acid code. Differences in translation between the germ line and rearranged sequences is also shown above the respective V gene. Junctional variation in the exact joining point between elements length can be seen. Non-germ-line N-region sequences at the joining sites are labeled. Random nucleotide differences within the body of an element are the result of somatic hypermutation.

the Ig sequences). As regards the first possibility, it should be noted that the heptamer sequences of TcR gene segments are not as conserved as those of Ig (data not shown). Rearrangement efficiency can apparently be influenced in B cells by differences in the heptamer sequence (Goodhardt *et al.*, 1987).

A second process adds random nucleotides between gene segments during joining (Fig. 6) (Alt and Baltimore, 1982; Kurosawa and Tonegawa, 1982). These "N regions" can range from 1 to greater than 10 nucleotides and are found in V_H (see Tonegawa, 1983) and in all four characterized T-cell-expressed chains (see Kronenberg *et al.*, 1986; Quartermous *et al.*, 1986; Chien *et al.*, 1987a). Human and mouse V_L sequences do not appear to include N regions, but rabbit V_χ chains may (Heidmann and Rougeon, 1983). Interestingly, mouse χ constructs transfected into rearranging B cells can include the N sequence upon rearrangement (Lewis *et al.*, 1985). This certainly implies that light chain sequences can be productive substrates for the enzyme(s) mediating N-region diversification. It is possible that the lack of recognized N regions in mouse and human V genes reflects a biased sample set, rather than biological constraints. All known Ig N sequences have a high G/C content. However, this bias is considerably reduced in mouse and human V_β N regions (data not shown). Potential V_α N sequences do not appear to be G/C rich either, but the lack of germ line data and the consequent difficulty in rigorously defining the germ line contribution to V_β and V_α junctional regions make the calculations less than certain. The V_δ N sequences, on the other hand, have a clear G/C bias. A G/C bias reduces the possibility of generating stop codons as well as several amino acid codons, such as those for aromatic residues. It is possible that clonal selection favoring or not favoring G/C-rich codons may explain this discrepancy among chains, rather than specific differences in the diversifying processes. Selection at both the DNA and protein level may play a role in this bias (Milner *et al.*, 1986). It has been suggested that the N sequence is the result of the operation of deoxynucleotidal transferase (TdT) on the free ends of DNA exposed during V gene rearrangement (Alt and Baltimore, 1982). TdT is known to add a random sequence with a G/C bias to a free end of DNA (Kunkel *et al.*, 1986), such as might be expected to exist during rearrangement. A correlation has been demonstrated between the extent of N-region sequence and the expressed levels of TdT (Desiderio *et al.*, 1984).

Both junctional and N-region diversification alter the third HV loop of Ig and TcR V regions and can have a profound effect on the repertoire of antigen specificities (e.g., Auffray *et al.*, 1981; Azuma *et al.*, 1984; Darsley and Rees, 1985; Hedrick *et al.*, 1988). In fact, these mechanisms result in the greatest variability of each of these chains being concentrated

in their third HV regions, with the possible exception of V_L chains (Barth *et al.*, 1985). This variability can be so great that essentially no recognizable germ line D gene segment sequence remains (Fig. 7). The role of the D gene segment may be more as a substrate for the diversifying mechanisms than as a contributor of specific amino acid sequence information.

Ig genes undergo a third type of somatic variation not directly linked to rearrangement. This process, denoted somatic hypermutation (Fig. 6), distributes point mutations throughout both V_H and V_L genes and flanking sequences, resulting in as much as a 3% nucleotide change from the germ line sequence (Kim *et al.*, 1981). The rate of hypermutation may approach 1×10^{-3}/base/cell generation (e.g., Wabl *et al.*, 1985; Sablitzky *et al.*, 1985), which is three to four orders of magnitude greater than expected in most cell types (Kimura, 1983). Hypermutation may contribute less than 5% of the total sequence diversity of the expressed V_H chain repertoire of the mouse (Gojobori and Nei, 1986). However, as many as 90% of B cells may express V_H genes that have undergone somatic hypermutation. These mutations may be clustered and are localized to about 1 kb of sequence centered about the rearranged V gene (Kim *et al.*, 1981; Gearhart and Bogenhagen, 1983). Although it may occur in pre-B cells without surface Ig expression (Wabl *et al.*, 1985), hypermutation appears to be an ongoing process (Rudikoff *et al.*, 1984; Clarke *et al.*, 1985). Antigen selection appears to drive the preferential expansion of mutated clones with increased affinity, contributing to the affinity maturation observed in the development of a humoral immune response (Gearhart *et al.*, 1981; Clarke *et al.*, 1985; Sablitzky *et al.*, 1985). Somatic mutation can also generate new specificities (Diamond and Scharff, 1984; Giusti *et al.*, 1987). However, somatic hypermutation probably more often has impact on affinity maturation rather than on increasing the overall repertoire of antigen specificities. The V_α and V_β and apparently V_γ genes do not frequently, if at all, undergo hypermutation (Barth *et al.*, 1985; Ikuta *et al.*, 1985; Behlke *et al.*, 1985; Davis, 1985). Information on the V_δ genes is not yet available. It is possible that T cells do not diversify in this manner in order to avoid the potential for generating autoreactive specificities outside the selective environment of the thymus (Barth *et al.*, 1985; Honjo and Habu, 1985; Eisen, 1986). Consequently, it may be that TcR affinities for antigen will be on average lower than those of Ig. The possibility of generating autoreactive B cells may pose a smaller risk because of the regulatory safeguards provided by T cells.

C_H class switching is a further rearrangement process found in Ig genes (see Honjo, 1983). Initially, a rearranged V_H gene is immediately upstream of the first C_H gene, C_μ (Fig. 4) (Shimizu *et al.*, 1982), to which it splices to create a complete mRNA. During class switching,

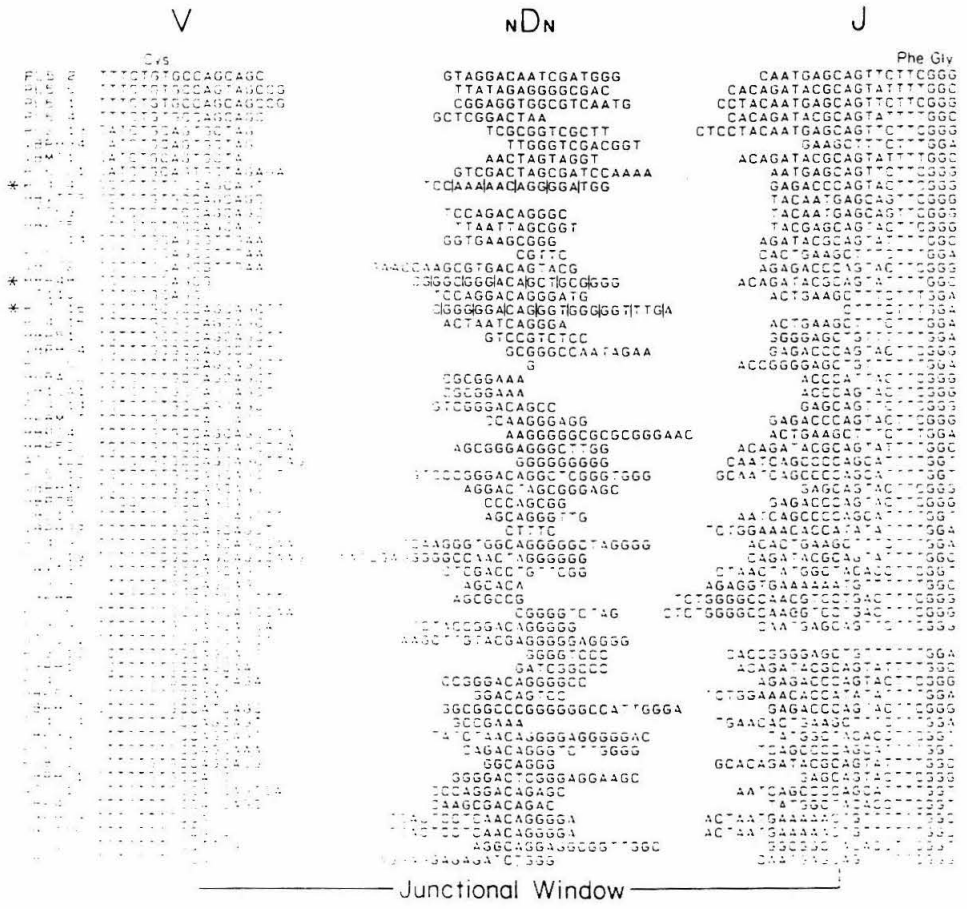


FIG. 7. Representative variation possible within a junctional window. The sequences in this figure are the 3′ sequences of a collection of human β-chain V genes (see Wilson *et al.*, 1988, for a key to the sequences). The V, D, and J gene segment sequences are labeled. The N-region sequence is included with the D sequences. An attempt was made to align the junctional sequences to the core D gene segments. However, so little of the germ line information remains in many of the regions that the alignment is often arbitrary. Likewise, the boundary between the V and junctional region is difficult to indicate precisely because of the lack of much germ line V gene segment information. The conserved Cys of the V region and the conserved Phe-Gly of the J segment are indicated for reference. Note the enormous variability of both length and coding potential between the 5′ most conserved J sequences and the 3′ most conserved V gene segment sequences even though only two D gene segments are used. Also, the (N)D(N) sequence can be translated in all three reading frames. The three sequences highlighted with an asterisk have the reading frame of their junctional sequence indicated by vertical bars separating codons. All three possible reading frames are represented.

the same V_H gene is juxtaposed 5′ to one of the other C_H genes further downstream within the C_H gene cluster, with the concomitant loss of the intervening DNA. The result is not diversification of the antigen recognition repertoire, but rather the expression of this diversity within the various functional contexts provided by the different C genes. This process occurs within repetitious "switch" sequences 5′ of most C_H genes. Like hypermutation, switching is often associated with the maturation of an immune response (Gearhart *et al.*, 1981).

C. Strategies of Diversification

There are obvious similarities among the gene families of the immune receptors (Fig. 4). However, each family and each receptor has a unique combination of strategies for organization and diversification, both species and chain specific. The λ L chain family is illustrative in this regard. The λ L chains are used in only 2-3% of mouse and rabbit Ig, but may comprise as many as 40% of the L chains of human and 95% of chicken Ig (Hood *et al.*, 1967). The mouse λ chain employs primarily three functional V/J pairs and is mostly diversified by somatic hypermutation (see Eisen and Reilly, 1985). The human λ family, though organized similarly to that of the mouse, has many more V and J elements and employs all of the general diversifying strategies discussed above to some degree. The chicken λ locus, on the other hand, has only a single functional V and J gene segment (Reynaud *et al.*, 1985), but has 25 tightly arrayed pseudo-V gene segments just 5′ to the functional V_λ gene segment (Weill, 1986). It appears that somatic diversification of the chicken λ repertoire results almost entirely through gene conversion-like events of the functional V gene segment by the pseudogene segments. Thus, a single V_λ gene segment appears capable of generating virtually the entire L chain repertoire of the chicken. The λ chain diversity in different species can obviously arise by very different mechanisms. Presumably then, the contribution of any gene family to the total repertoire of Ig or TcR diversity is determined by its overall diversifying potential rather than merely by its size.

The number of gene segments employed by various gene families is illustrative of how combinatorial potentials can vary in a family and in a species-specific manner. For example, the mouse has >500 V_H gene segments (Livant *et al.*, 1986), but <30 V_β gene segments (Patten *et al.*, 1984; Barth *et al.*, 1985; Behlke *et al.*, 1985). The human, on the other hand, may have only about 50 V_H gene segments but 75 V_β gene segments. Both the mouse and human may have less than 100 V_α gene segments and no D_α gene segments. However, the α locus has 5 to 10 times more J gene segments than other loci (Arden *et al.*, 1985; Becker *et al.*, 1985; Hayday *et al.*, 1985a; Winoto *et al.*, 1985; Yoshikai *et al.*, 1985). As Table I illustrates, when all variations are considered,

mouse B cells and T cells are probably capable of generating comparable levels of combinatorial diversity (Goverman *et al.*, 1986). Presumably, the same is true for other mammals as well.

As great as the combinatorial diversification potential may seem, the gene segment combinatorial values of Table I imply a much more modest diversity potential for Ig and TcR than would be expected if all diversifying mechanisms are factored in. It is impossible to be rigorous in quantifying the contributions of these mechanisms, but reasonable approximations can be made. Basically, instead of simply factoring in D segments in the combinatorial calculation, it may be better to quantify the total number of potential junctional sequences that can be generated between V and J gene segments, the loop of the third hypervariable region. The sequence in this region reflects the effects of N-region diversification, joining-site variation, multiple D elements (translated in three reading frames), and nonstandard rearrangements. Diversification from these mechanisms can be so great that each V gene family has a certain junctional window that can be virtually any sequence at the protein level. This window is defined as the sequence between the most C-terminal V segment residue that is never altered from germ line by junctional variation and the equivalent most N-terminal residue of the J regions (Fig. 7). Each chain has a characteristic minimum and maximum length for this window. Examination of known Ig and TcR HV3 regions suggests that all 20 amino acids can be found at each residue position in this region. Therefore, to a first approximation the number of potential junctional sequences for any chain is

$$\sum_{N\,=\,\min}^{\max} 20^N$$

where max is the maximum length of the window and min is the minimum length. The window for known V_β sequences for example is from about 6 to 15 amino acid residues. In practice, however, not all lengths within this range are equally represented, as this distance is only seldom greater than 8 residues. The similarly limited window size of V_α sequences appears to be from 3 to 7 residues. The number of sequences between 6 and 8 residues in length with any amino acid possible at any position is almost 2.7×10^{10}. When this value (and 1.3×10^9 for V_α) is factored into the combinatorial "V × J × junctional sequences" (using values from Table I), 4.4×10^{13} V_β and 6.5×10^{12} V_α chains can be generated. Assuming that only 1% of each of these is a viable protein because of inherent constraints of protein structure (e.g., chains with eight cysteines in a row are not likely to be functional), there is still the potential for generating 2.9×10^{22} receptors. If the number of these

pairs is reduced by another factor of 100 because of the inability of certain dimers to form or removal because of self-reactivity, 2.9×10^{20} potential TcR are possible from the germ line repertoire of one mouse. During the lifetime of a mouse, probably fewer than 10^9 thymocytes will ever leave the thymus, perhaps 1% of all those actually generated there (E. Rothenberg, personal communication). The difference in the potential TcR diversity and the actual level of diversity manifested in T cells is striking, particularly considering that these estimates of potential are certainly conservative. It is interesting that poorly formed chains and/or dimers may be more responsible for elimination of maturing thymocytes before migration to the periphery than selection against autoreactive clones. This could imply a role in T cell development for a test for the functionality of the expressed TcR gene products as a major aspect of MHC restriction, rather than a test for self-recognition per se (Goverman *et al.*, 1986).

Similar calculations for Ig (a window of from 5 to 10 for V_H regions and only 2 to 4 for V_L regions) (Chothia and Lesk, 1987) would result in 3.9×10^{18} potential Ig molecules. This may seem substantially less than the number of potential TcR, but Ig undergoes the added diversification of somatic hypermutation. Calculating the effects of hypermutation is difficult, but assuming that the product of any B cell division is susceptible to the process, the potential number of unique clones must be several orders of magnitude greater than the combinatorial potential alone. Again, compare these figures to a possible 10^{11}–10^{12} total B cells generated over the lifetime of a mouse.

Calculations of the potential diversity of γ/δ TcR are even less certain because the number of V_α gene segments that may rearrange with C_δ is unclear. However, to visualize the potential combinatorials it is sufficient to merely calculate the junctional sequence diversity potential of the V_δ sequences known (Elliot *et al.*, 1988). A window length of between 5 and 16 is readily possible, much wider than any other V family and primarily the result of the frequency of D_δ–D_δ joining. A window this size could potentially generate 6.9×10^{20} V_δ junctional sequences. Though it is certain that not all of these sequences are viable and that there will be skewed distribution of certain amino acids (because of the contribution of the D segments), the true diversity potential for a γ/δ dimer must be enormous. [Using a different method, Elliot *et al.* (1988) estimate a potential diversity of 10^{17} γ/δ dimers.] The potential is amazing when it is considered that there may be fewer than 10 V_γ or V_δ gene segments.

Of course, the diversity potentials just calculated primarily reside in the HV3, critical for determining the specificity of both Ig and TcR. It seems, in fact, that the presence of D segments and/or expanded J gene

segment families is the means by which to promote the complete diversification of the junctional region and can be considered as essentially equivalent diversifying strategies. However, HV3 is only one of three or four regions of each chain that contribute most to antigen specificity. Also, even residues outside the HV regions can interact with the antigen on occasion (e.g., Sheriff *et al.*, 1987). It is possible that alteration of HV3 often adjusts the specificity of a receptor that has a more general, but still restricted specificity provided by the V segment structure. This can be seen in fine specificity differences between TcR sequences that differ only in the HV3 (Winoto *et al.*, 1986; Fink *et al.*, 1986). In other words, it is selectively important to provide not only a large number of different receptors, but to ensure as broad a range of differences between these receptors as possible. Even though there are plenty of examples of single residues having profound effects on specificity, the most reasonable assumption is that the more nearly similar two receptor sequences are, the more nearly their specificities will overlap. No matter how diverse a junctional window is, if the rest of the sequence is essentially the same as many of the other members of the V gene family, the overall range of specificities will be severely truncated. Therefore, a diverse germ line repertoire of V gene segments may have more to do with providing this range than in contributing to greater combinatorial numbers. Certainly, from our previous calculations the number of potential receptors would not be much affected relative to actual clones expressed even if the number of V segments is reduced substantially (compare V_H and V_β, for example). In Table I there is a value for the number of known subfamilies of each chain. Subfamilies are defined by members of a particular gene family that are 75% or more similar (that cross-hybridize). The number of subfamilies, therefore, is one indication of the range of differences within a family (Barth *et al.*, 1985). The V_β and V_α gene families, although they have generally smaller subfamilies than Ig V gene segment families, have many more of them relative to the total number of genes and, hence, have much greater variance by this criterion than either V_H or V_L families. [Even though the maximum difference reported between V_β gene segments and between V_H gene segments is similar (Barth *et al.*, 1985), if two randomly picked V_H gene segments are selected, they will by chance be more similar to each other than two randomly chosen V_β sequences.] Does this imply a greater need for diverse TcR than Ig? Again, it must be remembered that hypermutation diversifies the entire Ig V region, perhaps mitigating any differences in range. The generally greater number of Ig V segments should also be considered. Also, experimental sampling differences could give a false impression about the true extent of germ line diversity in these families (Barth *et al.*, 1985).

The use of such varied diversifying strategies by the different gene

families and species argues that it is the total receptor diversity and the range of that diversity that generally represents the selectable phenotype, rather than particular sequences or modes of variation. This is an important perspective when considering the functional implications of the different strategies and their evolution. Most importantly, the level of potential diversity implies that virtually every pre-B and pre-T cell generates a unique receptor structure. Given even conservative estimates, the potential number of different receptors is certainly much greater than the number of different B or T cell clonal types expressed in any individual at any given moment (10^6–10^7; Jerne, 1955; Klinman *et al.*, 1976). The potential is so great for Ig that the loss of large segments of the germ line contribution can fail to have an appreciable impact on the level of functional diversity (Vice *et al.*, 1970; Kelus and Weiss, 1977). This variation is uniquely accommodated by a basic receptor structure that essentially allows the somatic tailoring of immune receptor repertoires to the current antigenic environment.

IV. Gene Organization and Diversity of the Major Histocompatibility Complex

Class I and class II MHC molecules are both noncovalently linked heterodimers (Fig. 1). Class I molecules are constructed of a class I heavy or α chain and a non-MHC-encoded light chain, β_2-microglobulin (β_2-m). Class II molecules consist of α and β chains of similar size. Both class I and class II molecules appear to fold into paired membrane-distal and -proximal domain structures (Fig. 1). The proximal domain sequences of all four chains are composed of classical C homology units (see Hood *et al.*, 1983). The distal regions of both class I and class II molecules have totally distinct structures. Therefore, the possibility of an evolutionary relationship between the distal regions and Ig homology units must remain conjecture (Ohnishi, 1984).

It seems that MHC molecules act as low-affinity, broadly specific receptors that interact with and present short peptide antigens to TcR (see Claverie and Kourilsky, 1987). These interactions are mediated by the distal domains (see Germain and Malissen, 1986) and hypervariable sequences analogous to those of Ig (see Mengle-Gaw and McDevitt, 1985). Although not generally similar, the distal sequences of class I and class II molecules do share specific sequence and structural similarities that are probably homologous in origin (Malissen *et al.*, 1984), implying not only a direct evolutionary relationship, but also suggesting that they present antigen in a similar fashion (Germain, 1986). The presented peptides arise as products of an intracellular "processing" or degradation of protein antigens. Processed peptides have been shown to bind to a single site on class II molecules (Giuillet *et al.*, 1986).

Recently, a high-resolution crystallographic analysis has been completed on a human class I molecule (Bjorkman *et al.*, 1987a). This structural data confirm that the membrane-proximal region of the α chain and β_2-microglobulin are canonical Ig C homology units. However, the most striking feature of the structure is the nature of the antigen-presenting, distal regions. Each of the two regions consists of four antiparallel β strands in a single sheet, overlayed by an extended α-helical structure. Interdomain alignment creates essentially a single, eight-stranded β sheet that appears as a table topped with two parallel helical structures running its length, forming a groove between them (Fig. 8). The residues lining

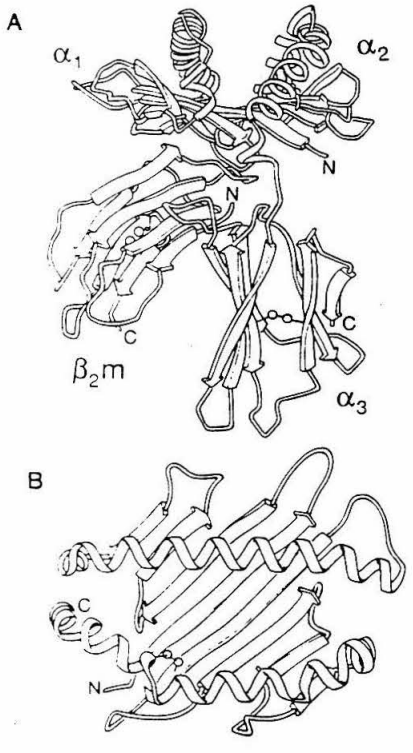


FIG. 8. Tertiary organization of an MHC class I protein. This figure (after Bjorkman *et al.*, 1987a) illustrates two views of a class I MHC molecule. A full-length side view (A) shows all three regions (α1, α2, and α3) of the class I molecules as well as their relationship to β_2-microglobulin. The β strands are shown as flattened arrows. The two long α-helical structures resting on the β-sheet platform of the α1 and α2 regions and the groove between them are viewed end-on. A view directly above the α1/α2 β-sheet platform (B) looks directly into the antigen-presenting groove.

this groove as well as those on the "top" surface of the helices contain most of the previously identified hypervariable residues of class I sequences (Cowan *et al.*, 1987; Gussow *et al.*, 1987). Therefore, it appears that the helical groove is the binding site of the peptide antigens (Bjorkman *et al.*, 1987b).

In various strains of mice, there are from 25 to 36 class I genes, but only two or three (K, D/L) are generally involved in antigen presentation and are coexpressed on most somatic cells (see Hood *et al.*, 1983). There are two class II loci, I-E and I-A, each encoding an α and β chain (see Mengle-Gaw and McDevitt, 1985). The class II molecules of both loci (isotypes) are coexpressed, but usually only on specialized antigen-presenting cells. The rat and human MHC are similarly organized (see Howard, 1987; Wake, 1986). Class I and class II loci involved in antigen presentation are the most polymorphic protein coding loci known in vertebrates. At least 50–100 alleles of each expressed locus are found in both mouse and human populations, and none of these loci is dominant (see Klein and Figueroa, 1981). Analogous to the V/C dichotomy of Ig and TcR, most of the heterogeneity is found in the membrane distal regions and can reflect a variation of up to 10% between allelic proteins (see Mengle-Gaw and McDevitt, 1985; Klein and Figueroa, 1986). Although there are interesting exceptions (Streilein and Duncan, 1983; O'Brien *et al.*, 1985), most mammals studied appear to share this high level of MHC polymorphism. This heterogeneity and multigenic expression assure the presence of multiple class I and class II molecules in normal individuals. For example, 75% of Caucasians will express four different alleles at their three class I loci (Bodmer and Bodmer, 1978). Because of this heterozygosity, class II α and β proteins can contribute further to this diversity in an intraisotypic combinatorial manner analogous to that of the two chains of Ig or TcR (see Bevan, 1981). However, limitations on particular intraisotypic pairings have been seen and may account for the tight genetic linkage between pairs of α and β class II genes (Braunstein and Germain, 1987). Combinatorial products between the different class II isotypes have also been reported (Germain and Quill, 1986; Malissen *et al.*, 1986), but the frequency of functional interisotypic combinations is unknown.

MHC polymorphism and combinatorial associations directly complement the TcR repertoire by expanding the context for antigen presentation, thereby broadening the range of antigen recognition possible with a given set of TcR (Goverman *et al.*, 1986). Unlike Ig and TcR, MHC molecules must rely solely on evolutionarily acquired variation. There are two recombinational mechanisms of diversification operating on MHC genes as a result of their multigenic organization. The first is the

process of homologous, but unequal crossing over. Such recombination events lead to the expansion and contraction of multigene families and changes in the repertoire of variant members (see Hood *et al.*, 1975). The variation in the gene number of both class I and class II genes seen among mouse strains is consistent with a frequent occurrence of such events (Rogers, 1985; Stephan *et al.*, 1986). This mechanism can also generate functional hybrid variants if recombination occurs within the genes (Sun *et al.*, 1985).

"Minigene" conversion is a second, presumably recombinatorial, mechanism affecting MHC gene families (Mellor *et al.*, 1983; Pease *et al.*, 1983; Weiss *et al.*, 1983). Independent events often generate the same clustered set of point substitutions (Geliebter *et al.*, 1986) and appear to result from small, gene conversion-like events between similar sequences. These events may account for much of the high rate of mutation at some MHC loci, as high as 10^{-3}/gene/generation in mice (Melvold and Kohn, 1975). The donor sequences of analyzed examples appear to be the non-antigen-presenting class I genes of the Qa and Tla loci (Mellor *et al.*, 1983). Interestingly, certain K alleles appear to undergo spontaneous gene conversion-like events an order of magnitude more frequently than the D alleles. Thus, position in the multigene family may dictate the frequency of recombination-like events.

While miniconversion events will extend diversity between alleles, they will also maintain a higher level of similarity between gene family members in general (a conversion inherently makes the two involved sequences more similar). Therefore, while frequent minigene conversions will maintain a dynamic heterogeneity between alleles, it will also provide a force for concerted or coincidental evolution within a population, particularly if the donor and recipient genes are not extremely tightly linked and/or the correction is directional in nature (Dover, 1982). Such conversion and homologous recombination events are probably responsible for the high degree of species-specific character found in MHC gene families (Hood *et al.*, 1975; Ohta, 1982; Bregegere, 1983; Hayashida and Miyata, 1983).

Once variants have arisen, what are the forces responsible for their fixation within a population? Are there positive selective forces favoring variants or is their establishment generally neutral? Is the process of variation itself selected? These are vexing questions because, even though there are no particularly frequent alleles, the collection of alleles is relatively stable within wild populations over time (Arden and Klein, 1982; Hayashida and Miyata, 1983; Figueroa *et al.*, 1985). Studies of variation between allelic polymorphic regions of MHC sequences demonstrate that the pattern of silent and replacement base changes

more nearly resembles that expected for random substitution than that of conserved sequences like globins (Hayashida and Miyata, 1983; Gustafsson *et al.*, 1984). These values imply that MHC polymorphic regions are under relatively weak constraining pressures as to their exact protein sequence. Paradoxically, the absolute rate of evolutionary change between class I alleles and species homologs is actually less than that of most other genes analyzed, including the globins (Hayashida and Miyata, 1983). This might indicate that structural variation is sometimes less important in the effects it can have on the interaction between an allele and antigen or TcR per se, than in how concomitant variation in its specificities complement or fail to complement the collective specificities of the allelic population. Thus, the lack of frequent MHC alleles implies that there is no "best" collection of MHC genes and that selection for variation occurs in the context of a population and particular lifestyle (Streilen and Duncan, 1983). This reflects the ability of pathogens to alter their antigenic profile more rapidly than host populations can acquire new MHC alleles. Because of its degenerate specificity, mutations leading to the improvement in the interaction of an allele with one antigen may weaken its interaction with other antigens, potentially exposing the organism to a host of other infectious agents. Therefore, within a changing antigen mileu no single MHC molecule or limited collection of molecules can ever optimally present all possible antigens. It is not only advantageous for an individual to be heterozygous for MHC molecules (in order for it to be able to respond to the broadest range of antigens), but also for it not to duplicate completely the repertoire of a neighbor. Pathogens have a much more difficult time expanding within a heterogeneous population in which the neighboring host organism will have a different repertoire of antigen-presenting alleles to be circumvented than the one already successfully invaded. The negative consequences of losing this populational diversity may have recently been demonstrated in cheetahs (O'Brien *et al.*, 1985). Cheetahs are much more susceptible to viral epizootics than are other big cats. It appears that cheetahs have undergone at least two severe populational bottlenecks (O'Brien *et al.*, 1987) that, because of founder effects, have essentially eliminated diversity of the MHC locus within the species.

The mechanisms responsible for assuring MHC variation are not special to the MHC. Rather, they are inherent consequences of the organization of the multigene families encoding the MHC. We assume that analogous recombination mechanisms also play significant roles in generating the germ line variation within the Ig and TcR gene families (e.g., Krawinkel *et al.*, 1983; Hayashida *et al.*, 1984; Clarke and Rudikoff, 1984; Kodaira *et al.*, 1986). However, Ig V gene segments do not appear

to undergo homologous crossing over as frequently as do MHC genes. This may reflect the lack of conserved linear sequences in V genes that would facilitate crossing over or correction and a shorter target length. Such factors would presumably also affect the rate of gene conversion and may account somewhat for the difference in conversion rates between K and D class I loci. Therefore, differences in the organization of MHC genes and V gene segments may affect not only the rate, but also the nature and mode of their evolutionary diversification.

For both MHC and V gene families, it appears that the overall repertoire of diversity, rather than the individual variant, is generally the selectable phenotype. However, it is presumably the clonal expression (one receptor per cell) of Ig and TcR diversity versus the nonclonal expression (multiple receptors on each cell) of MHC diversity that establishes the selective constraints on the gene copy number of each family and, hence, the differences in how diversity in each system is selected and maintained. For the large Ig and TcR gene families these organizational differences reflect a selection for diversity within a family rather than between alleles. On the other hand, a single antigen-presenting cell presents multiple peptide/MHC complexes at any one time. It is the combination of peptide/MHC molecule that is the functional T cell antigen. The same peptide in association with a different MHC molecule is effectively a different antigen and would generally be recognized only by different T cells. There is certainly a practical limit on how many MHC molecules per se can be expressed on a single cell as well as a limited amount of any particular processed peptide antigen available for presentation by that cell. Too many coexpressed class I or class II variants would lower the concentration of any particular peptide/MHC combination, potentially below the threshold level required to trigger a specific T cell. With such constraints, diversity of the smaller repertoire of MHC genes is assured through the maintenance of extreme allelism and a heterozygous population. This extreme allelism ensures that the broadest possible range of antigen specificities is available to every antigen-presenting cell in a manner analogous to the contribution to the range of T cell specificities by the large number of subfamilies in TcR V genes.

V. Nonimmune Receptor Members of the IgGSF

At least 23 other distinct genes or gene families with no direct role in antigen interaction have been identified as belonging to the IgGSF. Most are single-gene members and most appear to encode distinct cell surface, receptor molecules. Although these genes are generally nonpolymorphic, the diversity of the examples is striking. For convenience,

these molecules can be loosely associated in eight categories: (1) non-antigen-presenting, β_2-microglobulin-associated molecules, (2) T-cell-associated molecules, (3) molecules expressed on both T cells and nervous system cells, (4) nervous-system-associated molecules, (5) Ig-binding molecules, (6) growth factor/kinase receptors, (7) miscellaneous and (8) uncertain examples. Note that these categories are not meant to imply necessarily functional or evolutionary relationships.

A. Non-Antigen-Presenting, β_2-Microglobulin-Associated Molecules

As mentioned, β_2-microglobulin is a single C homology unit, the light chain of the MHC class I molecule (Cunningham *et al.*, 1973). It is probably divergently related to the MHC class II α chain (McNicholas *et al.*, 1983) and may be considered functionally an orphan MHC gene. It is encoded by a single, nonpolymorphic gene (Parnes and Seidman, 1982). Besides the antigen-presenting class I sequences, β_2-microglobulin is also found in association with the Qa and Tla MHC molecules and the CD1 family of antigens.

In the mouse, most MHC-linked class-I-like genes are encoded in the Qa and Tla regions and are not involved in antigen presentation. Although the Qa and Tla alleles are not as polymorphic as the K and D alleles, substantial rearrangements of the loci through unequal crossing over is seen (see Flavell *et al.*, 1986; S. Hunt, personal communication). These sequences are strongly conserved within species, but may not be conserved at all between species (Rogers, 1985). The function of putative Qa and Tla gene products is unknown, but it has been suggested that the genes serve primarily to drive the evolutionary diversification of the antigen-presenting K- and D-like alleles by providing a pool of diversity for gene conversion-like events (e.g., chicken λ; see Howard, 1987). However, the differential expression of many of these genes in distinct populations of hematopoietic cells and the relatively low frequency of obvious pseudogenes (S. Hunt, personal communication) argue that there may be selection of these sequences directly at the protein level, perhaps as elements involved in the differentiation of hematopoietic cell subsets. The chicken Vλ gene family indicates that there is no strict requirement for maintenance of open reading frames to be efficient donor sequences in gene conversion-like events. Accordingly, without selection at the protein level, one might expect a larger number of pseudogenes within this group. Also, some Qa molecules are secreted and may have regulatory functions (Soloski *et al.*, 1986).

The human CD1 antigens are differentiation antigens that characterize immature thymocytes and some lymphoid malignancies (Bernard *et al.*,

1984). Like class I sequences, CD1 molecules associate noncovalently with β_2-m (Ziegler and Milstein, 1979), but at least one CD1 antigen (T6) is also found disulfide linked, in some thymocytes, to the CD8 molecule (Ledbetter *et al.*, 1985; Snow *et al.*, 1985), another member of the IgGSF (see below). Recent characterization of the CD1 genes indicates that they have an exon and protein structure similar in organization to class I heavy chain genes, with one membrane-proximal, C homology unit and two unrelated distal regions. However, they are not linked to the MHC and they are no more similar to class I than to class II sequences (Calabi and Milstein, 1986). This family has at least five members that may be clustered. Hybridization results indicate homologous sequences are present in mice (Martin *et al.*, 1986). The role of these molecules is unknown, but their interspecies conservation, dissimilarity to MHC genes, and differential expression imply an independent protein function, possibly morphogenic in nature. The T6 molecule is expressed on all epidermal Langerhans cells. These are the only epidermal cells that express MHC class II molecules and are believed to be specialized antigen presenting cells (see Wolff and Stingl, 1983). Recent work reveals that the T6 molecule expressed on Langerhans cells is cointernalized with MHC class II molecules through the process of receptor-mediated endocytosis, indicating a potential role of the T6 antigen as a receptor involved in some immune function (Hanau *et al.*, 1987).

B. T-Cell-Associated Molecules

Besides the TcR, T cells express a host of accessory molecules that are presumably involved in signal transduction, cell adhesion, and even the facilitation of antigen/MHC targeting. The CD4 and CD8 molecules are accessory molecules of T cells that appear to play an important role in facilitating T cell interactions with target cells. The CD4 and CD8 molecules are expressed on most thymocytes together. However, most peripheral T cells fall into two mutually exclusive populations, those expressing only CD4 or only CD8 molecules. Both molecules appear to interact directly with either class I (CD8) or class II (CD4) molecules (Dembić *et al.*, 1987; Gabert *et al.*, 1987) and hence may be predominantly responsible for restricting cytoxic T cells to class-I-presented antigens and helper T cells to class-II-presented antigens (see Townsend, 1985). The human CD8 molecule is usually characterized as a homodimer of α chains, unlike the CD8 of mouse, which is a heterodimer with an α and β chain (Lyt-2 and L3T4, respectively). The CD4 molecule appears to be monomeric. However, recent evidence indicates that the human CD8 molecule may also exist as a heterodimer employing a Lyt-3 (β) homolog (J. Parnes, personal communication). Each molecule has an

N-terminal V homology unit, most like those of light chains, linked by a connecting sequence to transmembrane and cytoplasmic regions (Fig. 1) (Johnson *et al.*, 1985; Littman *et al.*, 1985; Maddon *et al.*, 1985; Sukhatme *et al.*, 1985; Zamoyska *et al.*, 1985; Johnson and Williams, 1986; Tourvieille *et al.*, 1986; Gorman *et al.*, 1987; Nakauchi *et al.*, 1987). Beyond these similarities, the CD4 and CD8 molecules have very different structures. The CD8 α and β chains have connecting sequences approximately 60 residues long. The CD8 α and, perhaps, the CD8 β chain have an immunoglobulin hingelike sequence in this region that presumably plays an important role in determining how the molecule is displayed into the extracellular environment (Sukhatme *et al.*, 1985). The CD8 β chain has a short peptide sequence that is indistinguishable from those of Ig and TcR J segment sequences (Fig. 9). Neither the CD8 α or β chain gene rearranges during differentiation. Considering the high level of identity between the CD8 β J-like sequence and those involved in rearrangement, this lack of rearrangement in CD8 β raises interesting evolutionary possibilities that will be discussed later. The CD8 α and β genes are single copy, are located within a few kilobases of each other, and are closely linked to the κ light chain locus (J. Parnes, personal communication). The CD8 homodimers and heterodimers appear to form single V domain-like receptors possibly interacting with a conserved epitope on MHC class I molecules (see Goverman *et al.*, 1986).

Even though CD4 and CD8 molecules appear to perform analogous functions with homologous ligands, class II and class I MHC molecules, respectively, they do not appear to share a recent common origin. The CD8 α and β V-like regions are more closely related to V_L sequences than the N-terminal V-like element of the CD4 molecule (Tourvieille *et al.*, 1986). Moreover, the CD4 molecule has a connecting sequence approximately 280 residues in length. Within this sequence and immediately C terminal to the first V-like sequence is a second, albeit truncated, homology unit sequence (Tourvieille *et al.*, 1986). It has

```
mouse Jλ con   WVFGGGTKLTVL
mouse Jκ con   .T.....EIK
mouse JH con   D.W.Q..TV..S
mouse γt3      M...T.....V
human CD7      N.Y.S..-..LV
rat MRC Ox-2   NM..S.-.VSGT
```

FIG. 9. The J-like sequences in nonrearranging IgGSF members. Shown for comparison are consensus J sequences for the three Ig families. Similarity to the J λ consensus is indicated by dots. Gaps inserted to optimize alignment between J elements are indicated with dashes.

been suggested that there are two even more extensively diverged homology units tandemly arrayed in the connecting sequence (Clark *et al.*, 1987), one without a disulfide loop. The second homology unit is more nearly similar to the first V-like region than to other IgGSF member sequences, indicating its probable origin through a duplication event involving the first V-like or related region (Tourvieille *et al.*, 1986). Therefore, although the foreshortened nature of the second homology unit might suggest that it belongs to the H class of sequences, its similarity to the V-like region implies that it should also be classified as V-like. The evolutionary implications of its truncated nature will be discussed later.

With such dramatic differences in structure, how do the the CD4 and CD8 molecules perform such apparently analogous functions? How can the CD4 molecule form a domainlike receptor structure like the CD8 chains presumably do? It is still possible that CD4 is expressed as a homodimer or heterodimer. As mentioned, it is only recently that a human CD β chain may have been identified, and a loosely coupled CD4 homodimer would be difficult to identify. Alternatively, two of the V-like units of the CD4 molecule, presumably the most N-terminal pair, could fold together to generate a pseudodimeric structure (Parnes and Hunkapiller, 1987). Monoclonal antibody epitope mapping of the CD4 chain does indicate that these two regions are in close enough apposition to generate a composite epitope (B. Jameson, personal communication). It is possible, however, that the CD4 and CD8 molecules perform their roles in completely unrelated ways. CD4 may interact monomerically with its MHC ligand through β-sheet interactions between a homology unit of CD4 and a C unit of the class II molecule in the manner of other IgGSF domain-forming interactions.

The α/β and γ/δ heterodimers are noncovalently associated on the T cell surface with the CD3 complex. The human and mouse CD3 molecules are composed of γ, δ, ϵ, and ζ chains. CD3 γ, δ, and ϵ cDNA sequences have been characterized (van den Elsen *et al.*, 1984; Krissansen *et al.*, 1986; Gold *et al.*, 1986, 1987). Each has a single H-type homology unit, a short connecting sequence, a transmembrane element, and a hydrophylic cytoplasmic tail. It is believed that an unusual asparagine residue conserved in the transmembrane region of all three chains is involved in forming a salt link with an equally conserved lysine residue in the transmembrane peptide of all four TcR chains. The γ and δ chains are each significantly more similar to each other than either is to any other IgGSF member, indicating a relatively direct evolutionary relationship (Krissansen *et al.*, 1986; Gold *et al.*, 1987). The CD3 ϵ chain, on the other hand, is not any more similar to the γ or δ chain than it is to several other members of the IgGSF, indicating a more distant

divergence. A particular relationship between the CD3 γ, δ, and ϵ chains and another H-structure IgGSF member, the neuronal-cell adhesion molecule (see below), has been argued on the basis of statistically greater similarity scores and close chromosomal linkage, each to band q23 of human chromosome 11 (Gold *et al.*, 1987). However, the scores appear more to indicate that these molecules share homologous H-structures rather than a particularly close evolutionary relationship. The Thy-1 gene maps as well to band q23, but has no particular similarity to the CD3 gene.

The CD28 (Tp44) molecule is a disulfide-linked homodimer expressed on $CD3^+$ T cells that coexpress α/β TcR (Hara *et al.*, 1985), but apparently not on peripheral T cells that express the γ/δ TcR (Poggi *et al.*, 1987). Monoclonal antibodies to the CD28 chain added in the presence of phorbol esters can cause $CD28^+$ T cell proliferation (Hara *et al.*, 1985). The normal ligand, and hence the function of the CD28 molecule, is still unknown. Its apparent absence from peripheral γ/δ TcR, however, suggests that it may play a role in as-yet undefined activation requirements of α/β T cells that might define functional distinctions between α/β and γ/δ T cells (Poggi *et al.*, 1987). The CD28 chain is encoded by a single-copy gene (Arrufo and Seed, 1987a) and has a single N-terminal V-like homology unit, a short connecting peptide, a transmembrane region, and a short cytoplasmic sequence. Immediately extracellular (N-terminal) to the transmembrane-spanning sequence is a highly distinctive hingelike region. A CD28 molecule is, consequently, organized very similarly to the CD8 molecule. It will be interesting to see whether the CD28 molecule plays an analogous accessory function.

The CTLA-4 gene was identified by analyzing a cDNA isolated from a subtracted mouse cytotoxic T-cell-derived cDNA library (Brunet *et al.*, 1987). There is a single CTLA-4 gene. The protein has not yet been identified, but the predicted protein sequence clearly suggests it is an IgGSF member. Like the CD28 antigen, the CTLA-4 protein has a rather classic V-like homology unit except for the absence of the V-region-invariant tryptophan (Fig. 3). In fact, when the CD28 and CTLA-4 sequences are compared, it is evident that they are very closely related (data not shown). The connecting sequence is particularly conserved (one stretch of 25 amino acids has 15 exact matches and 4 highly conservative substitutions). Also, all five extracellular cysteine residues are precisely conserved. It is quite possible that the CTLA-4 gene is the mouse homolog of the human CD28 antigen gene. Interestingly, the hingelike sequence in the CD28 chain is precisely deleted from the CTLA-4 chain. If the CD28 and the CTLA-4 genes are homologs, the absence of the hinge region in the CTLA-4 mRNA may reflect the deletion of an exon encoding this region during evolution or alternate RNA splicing.

The human CD7 (gp40) antigen is a highly glycosylated surface antigen of thymocytes and T cells (Palker *et al.*, 1985). It is one of the earliest and most definitive markers of cells belonging to the T cell lineage. Its function is unknown. It appears to be a single, 40-kDa chain. A CD7 cDNA sequence (Aruffo and Seed, 1987b) reveals an organization remarkably similar to that of the CD28 and CTLA-4 cDNAs, but it is not directly related to either by sequence similarity. Like the CD28 and CD8 α chains, the CD7 antigen has a pronounced hinge sequence just N-terminal to the probable transmembrane region. Also, like the CD28 and CTLA-4 chains, the CD7 chain has a classical V-like homology unit without the conserved tryptophan. The CD7 protein also shares an interesting feature with the CD8 β chain, an apparent J-region-like sequence located appropriately C-terminal to the V-like region (Fig. 9). Although the similarity of the CD7 J sequence to Ig or TcR J elements is not as dramatic as that of the CD8 β chain, it nevertheless maintains the core residues of the canonical J motif: a hydrophobic residue, an aromatic residue, a Gly, a small residue, a Gly, a Thr, any residue, and a hydrophobic residue (Fig. 9). In fact, the CD7 J sequence is more similar to that of the CD8 β chain than to any other J element: 7 of 11 residues are identical with one single residue gap and two highly conservative substitutions (Phe → Tyr, Thr → Ser). The CD7 gene, like the CD8 β gene, does not rearrange during differentiation (Aruffo and Seed, 1987b).

It seems that T cells express a whole class of accessory molecules with analogous, if not directly homologous, structures: one N-terminal V-like homology unit, a connecting sequence of approximately 50 residues, a proline-rich, hingelike sequence immediately N-terminal to a transmembrane sequence, and a relatively short cytoplasmic sequence 15–40 residues long, which is unrelated to the cytoplasmic regions of the other sequences of the IgGSF. The CD8 α and β, CD28, CTLA-4, and CD7 chains all belong to this group. Besides the hingelike sequence, each of these chains also has a conserved a cysteine residue just N-terminal to the transmembrane sequence. Cysteine residues associated with the hinge regions of immunoglobulins are involved in interchain disulfide bonds, as they appear to be in the CD8 and CD28 chains. The structural homology of these molecules suggests that they are displayed in a similar manner on the cell surface and probably interact with their ligands in a similar manner. If the CD8 molecule is typical of this group we would predict that the ligands for the CD7, CD28, and CTLA-4 proteins are superfamily members, perhaps even MHC or CD1 molecules.

C. T-Cell- and Nervous-System-Associated Molecules

The Thy-1 molecule is one of the simplest members of the IgGSF, with only a single V-like homology unit (Cohen *et al.*, 1981) attached to the

cell membrane by a glycophospholipid anchor (Tse *et al.*, 1985; Fig. 1). The murine Thy-1 molecule is found in abundance on thymocytes and neurons as well as fibroblasts and a variety of other cells. The function of Thy-1 is unknown. However, it apparently can act as a signal transduction molecule in cellular activation (Kroczek *et al.*, 1986). Although Thy-1 is generally isolated as a monomer, there is evidence that it can form multimeric complexes, probably through homophylic interactions (Xu *et al.*, 1987). This might imply a role for Thy-1 in cell-cell interactions mediated by homophylic interactions of Thy-1 molecules between cells.

The rat Ox-2 cell-surface antigen has a tissue distribution similar to Thy-1 (Barclay, 1981; Webb and Barclay, 1984) and likewise has an unknown function. Its structure resembles that of a TcR α or β chain, with V, C, transmembrane, and cytoplasmic regions (Clark *et al.*, 1985). Like the CD7 chain, the Ox-2 chain has a J-like region between the V and C regions (Fig. 9) (Barclay *et al.*, 1986). Ox-2 is presumed to be a single-chain molecule. However, its striking organizational similarity to other chains found only as dimers suggests that it is best to leave this question open.

The CD4 gene is also expressed in the brain, but the cell type(s) in which it is expressed and its functions are unknown (Tourvieille *et al.*, 1986). It is interesting that the predominant CD4 transcript in mouse brain encodes a protein with the first 200 or so N-terminal residues truncated, including the first two V-like regions (Gorman *et al.*, 1987).

D. Nervous System Molecules

The most intriguing new group of Ig-like sequences may be those expressed predominantly by cells of the nervous system. The first to be described was a neural cell adhesion molecule (N-CAM) isolated from chick embryo brain (Hemperly *et al.*, 1986a). The N-CAM gene encodes five H-type, N-terminal homology units, a long connecting sequence, a transmembrane region, and a very large cytoplasmic domain (Cunningham *et al.*, 1987; Barthels *et al.*, 1987). N-CAM belongs to a limited number of cell adhesions whose binding is involved in cell-cell interactions that coordinate patterns of migration, proliferation, and cellular differentiation during tissue development (see Edelman, 1986). N-CAM is found in several developmentally regulated and tissue-specific forms, apparently the products of alternate splicing pathways that include varying amounts of the cytoplasmic tail and determine whether membrane attachment is mediated via a transmembrane peptide sequence or a glycophospholipid anchor (Hemperly *et al.*, 1986b). These differences, along with modulation of the amount and distribution of

expression, correlate with different patterns of tissue formation. N-CAM binding is homophylic and polyvalent (Edelman, 1983) and maps to the Ig-like regions (Cunningham *et al.*, 1983). It is possible that these properties reflect Ig domainlike association between N-CAM molecules on different cells (Hunkapiller and Hood, 1986) and could have implications for other examples of interactions between other members of the IgGSF, such as the Thy-1 molecule.

Another brain protein, rat myelin-associated glycoprotein (MAG), is most similar in structure and sequence to N-CAM (Fig. 1) (Hunkapiller and Hood, 1986; Arquint *et al.*, 1987; Noronha *et al.*, 1987). MAG has two well-defined and possibly three additional relic H homology units as well as a transmembrane and large cytoplasmic region. It appears that MAG is involved in neuron–glial and glial–glial cell interactions mediating myelination and the maintenance of periaxonal space (Sternberger *et al.*, 1979; Nobile-Orazio *et al.*, 1984), possibly by providing an adhesion system in the manner of N-CAM (Riopelle *et al.*, 1986). MAG is found in low abundance in both central and peripheral nervous system myelin (Quarles, 1984). A protein has recently been described on both T and B cells as having serological and biochemical characteristics similar to those of MAG (Peault *et al.*, 1987). The protein is conserved phylogenetically and is regulated in a developmentally specific manner. The authors suggest that this protein is MAG and that it may be involved in the "homing" process of both immune and nervous system cells, mediated by its cell adhesion properties. It will be interesting to learn whether these putative homing interactions are regulated or facilitated by interactions between MAG and other immune/nervous system coexpressed members of the IgGSF such as the Thy-1 or Ox-2 antigens.

The similarity of N-CAM and MAG suggests there may be an entire family of IgGSF sequences involved in cell-to-cell interaction or adhesion in neuronal morphogenesis and raises the possibility that other CAMs also belong to the IgGSF. However, recent characterization of liver (L)-CAM reveals it to be distinct from the IgGSF (Gallin *et al.*, 1987). It is interesting to note, however, that L-CAM and several related CAM sequences (Ringwald *et al.*, 1987; Nose *et al.*, 1987) share an analogous structural motif with N-CAM and MAG. The apparent adhesive properties of each molecule is mediated by N-terminal repeat units approximately 100 residues in length. The non-IgGSF CAM sequences each have three or four such repeats that are obviously the products of internal duplication, as appears to be the case for N-CAM and MAG. These repeats, as with H-type homology units, are predicted to have secondary structures dominated by six or seven alternating β strands (data

not shown; after Chou and Fasman, 1978). Further analysis might elucidate whether the homophylic association of these CAMs is mediated by β-sheet interactions as would be predicted for N-CAM and MAG.

P_0 is a myelin-associated protein that constitutes over 50% of peripheral nervous system myelin, but is not found at all in the central nervous system. P_0 has a single external V-like homology unit (Fig. 1; Lai *et al.*, 1987) that is interestingly similar to the V-like unit of the Ox-2 antigen, suggesting an independent evolution from N-CAM and MAG. P_0 molecules are thought to facilitate the compaction of peripheral myelin by linking adjacent lamallae of Schwann cell membrane (Braun, 1984). It is suggested that P_0 brings the external membrane surfaces of adjacent lamallae together through homophylic interactions between the extracytoplasmic domains of P_0 molecules expressed on opposing membranes (Lemke and Axel, 1985), presumably through domainlike β-sheet interactions. The cytoplasmic sequence of P_0 is highly positively charged. The intracytoplasmic membrane surfaces could be compacted through electrostatic interactions of the positive charges of the P_0 cytoplasmic tail with the negative charges of the head groups of acidic lipids of the opposing membrane surface.

E. Ig-Binding Molecules

The poly(Ig) receptor (p-IgR) shuttles polymeric IgM and IgA antibodies from the blood side to the serosal side of mucous membranes (Brandtzaeg, 1981; Kühn and Kraehenbuhl, 1981; Mostov *et al.*, 1980). The external portion of the p-IgR is released during the process and is known as the secretory component (SC). The SC molecule has five V homology units (Fig. 1) (Mostov *et al.*, 1984). Comparison of the individual units indicates that they are each more closely related to each other than to other IgGSF members. Hence, they are likely to be the product of a series of internal duplication events that resulted in the expansion of a single unit sequence. It is possible that two or more units of p-IgR form a pseudodimer domain structure that acts as an Ig-like receptor for the Ig heavy chain. Alternatively, individual p-IgR homology units may interact with Ig units through domainlike β-sheet interactions, similar to the interaction of Ig units with themselves. These interactions would presumably be between p-IgR and the C-terminal regions of the heavy chain. Disulfide bonds can be generated between p-IgR and some subclasses of IgA (Brandtzaeg, 1981).

Humoral Ig/antigen complexes are bound by various cell types through receptors to a portion of the C_H region of Ig molecules known as the Fc portion. This process helps mediate cooperation between the humoral and cellular immune pathways (Springer and Unkeless, 1984; Dickler,

1982). The Fc receptors (FcR) for the mouse γ isotype have been characterized (Ravetch *et al.*, 1986). Two similar genes and alternate RNA processing are responsible for at least three different tissue-specific FcRγ. Each has an external region composed mostly of two tandem homology units, a transmembrane region, and different cytoplasmic regions. The two homology units are more similar to each other than to any other Ig-like sequence, and hence are probably the result of an internal duplication. Both domains are H unit sequences. It remains to be seen whether any of the FcR for other Ig isotypes belong to an encompassing FcR multigene family. However, at least one FcR for Ig ϵ does not appear to belong to the IgGSF (Kikutani *et al.*, 1986). The polydomain structure of the FcRγ is similar to that of p-IgR and may indicate analogous functional interactions with their Ig ligands, even though they are clearly not directly evolutionarily related.

F. Growth Factor/Kinase IgGSF Members

The receptor for platelet-derived growth factor (PDGFR) and the *fms* transforming gene have clearly related sequences and share an intracytoplasmic region similarity to *src*-related tyrosine kinases (Yarden *et al.*, 1986). The c-fms protein is thought to be the receptor for macrophage colony stimulating factor (CSF-1R) (Coussens *et al.*, 1986) and hence, both PDGFR and CSF-1R appear to be responsible for promoting cell proliferation and differentiation. Both PDGFR and CSF-1R have three and perhaps five extracytoplasmic H-type homology units (Lai *et al.*, 1987). Thus it seems that an IgGSF receptor has been coupled to a kinase activity, perhaps through an exon shuffling event, and then diverged to generate a family of growth factor receptors. The similarity of the intracytoplasmic region of CSF-1R and that of a partial cDNA of the *v-kit* oncogene (Yarden *et al.*, 1986) suggests that *c-kit* also belongs to this family. Except for antigen receptors and link protein (see below), PDGFR and CSF-1R are unusual as examples of IgGSF receptors with characterized ligand that are not also IgGSF members.

G. Miscellaneous Members

Carcinoembryonic antigen (CEA) is the most commonly used human tumor diagnostic antigen (see Shiveley and Beatty, 1985). Although it is not strictly tumor specific, its expression is greatly increased in colonic tumors and possibly those of lung and breast. CEA represents an entire family of heterogeneous, highly glycosylated proteins, including several serologically related antigens; e.g., nonspecific cross-reacting antigen (NCA), biliary glycoprotein I, and normal fecal antigen (NFA-1). The distribution of expression of members of this family is extensive,

including some within the immune system, but no functions are known. Recent protein and cDNA sequences reveal that CEA is an integral membrane protein with six external Ig homology units (Paxton *et al.*, 1987; Thompson *et al.*, 1987; Zimmerman *et al.*, 1987). These are remarkable in that they represent three tandem repeats of two H-type units each. It appears there are about a dozen CEA-related genes by hybridization. The heterogeneous nature of the proteins suggests that varying numbers of tandem repeat units may also be found in different members. (While CEA has three, it appears that NCA-55 may contain only one repeat and NCA-95 contains two.) Each CEA-related protein also has an N-terminal sequence, with no cysteine pair, about the same length as the homology units. It is possibly an evolutionary relic of a homology unit (Paxton *et al.*, 1987).

The α1B-glycoprotein (α1B) is a prevalent human plasma protein of unknown function (Schultze *et al.*, 1963). Like SC, α1B has five N-terminal homology units (Ishioka *et al.*, 1986). It has been suggested, therefore, that α1B may be the secretory component of the p-IgR for IgG antibodies (Ishioka *et al.*, 1986). Also, like p-IgR, α1B appears to be the product of multiple internal duplication events. Unlike p-IgR, however, α1B is constructed of H homology units, like most of the other polydomain IgGSF members. An interesting feature of the α1B structure, which appears to distinguish it from other polydomain IgGSF members, is the presence of a proline-rich hingelike sequence between each of the homology units. How such structures would affect the quaternary relationship of the individual units is unclear. It is interesting to note that rigid hinge sequences producing similar angles between adjacent units, could generate a spiral or circular arrangement of homology units.

The proteoglycan aggregates of cartilage are composed of proteoglycan monomers associated with hyaluronic acid (HA). Link glycoprotein stabilizes the aggregate structure by binding to both proteoglycan and HA. These interactions result in the formation of an extracellular matrix of cartilage. A prominent species of link protein has been sequenced using material from a rat chondrosarcoma and bovine nasal cartilage (Neame *et al.*, 1986; Bonnet *et al.*, 1986). Surprisingly, the link protein was found to belong to the IgGSF. This may be the only known example of an IgGSF member that may never be deployed in the context of a membrane-bound receptor (as mentioned, α1B may be a secretory component of a membrane-bound molecule). Link protein is essentially divided into two functional regions. The N-terminal 125 amino acids represent a specialized V-like homology unit. It has the repertoire of V-like residues, but is longer between the two cysteines than any other

known member by about eight residues. This length difference can be accounted for by extra residues between the N-terminal conserved cysteine and the conserved tryptophan, the region that corresponds to the HV1 of V regions. The remaining 60% of the sequence is constructed of two internally homologous non-IgGSF regions. It is probably these two repeats that interact with HA, while the V-like region binds the proteoglycan (Neame *et al.*, 1986). The sequences of rat and bovine link proteins are extremely conserved, including the HV1-equivalent loop (one mismatch in 20 residues). This contrasts dramatically with the lower level of conservation seen across species for many of the other IgGSF receptors, such as CD8 and CD4 molecules. When the nature of the interactions between these molecules is better understood, it will be interesting to determine whether the extra loop sequence is involved in generating a specilized binding pocket. Developing reasonable scenarios for the evolutionary relationship of the link protein to other V-like sequences should also prove interesting.

H. Uncertain Members

It should be noted that several other sequences have been proposed as members of the IgGSF on the basis of limited stretches of identity with other members. Several of the polydomain members (i.e., CD4, CEA, PDGFR) have sequences that may be relic homology units. This possibility is supported by statistical analysis as well as their presence in molecules with other well-defined IgGSF domain sequences. Several of these relic regions do not have the conserved cysteine pair. However, there is at least one excellent example of a functional immunoglobulin V region also lacking the disulfide bridge (Rudikoff and Pumphrey, 1986). Thus, it is clear that no single rule of membership in the IgGSF is absolute.

Whether or not the sequences mentioned above are IgGSF homology unit descendants is irrelevant to defining the molecules themselves as members of IgGSF, because each does contain rather obvious homology units. However, there are a number of sequences where the history is a bit more ambiguous. The CD2 molecule is one example. The CD2 (T11) antigen is expressed on $CD4^+$ and/or $CD8^+$ thymocytes and T cells (Reinherz, 1985). Monoclonal antibodies to the CD2 antigen in the presence of phorbol esters can initiate T cell proliferation. The CD2 molecule is involved in antigen-independent adhesion of thymocytes to thymic epithelial cells and adhesion of cytotoxic T cells to target cells. Its ligand in these reactions is the LFA-3 molecule, a rather ubiquitous cell surface marker (Takai *et al.*, 1987). On the basis of limited stretches of similarity to various members of the IgGSF, it has been claimed that

the CD2 molecule is also a member (Sewell *et al.*, 1986). The recent characterization of LFA-3 (Wallner *et al.*, 1987; Seed, 1987) indicates that the CD2 and LFA-3 chains are homologs. It is suggested that the CD2 and LFA-3 molecules are constructed of an N-terminal homology unit without a disulfide bridge immediately followed by a disulfide-defined unit, a short connecting sequence, a transmembrane peptide, and a cytoplasmic region of moderate length. The LFA-3 chain can be membrane linked by a glycophospholipid anchor (Seed, 1987). Although statistical similarity scores between one region of the CD2 chain and various IgGSF members (primarily the CD4 molecule) have values >4 SD (Williams *et al.*, 1987), the sequence does not share the conserved cysteines or many of the other IgGSF conserved residues.

Blast-1 is an activation marker of B cells (Thorley-Lawson *et al.*, 1982) and probably T cells (Staunton and Thorley-Lawson, 1987) that forms a noncovalently linked heterodimer with another, uncharacterized protein (Thorley-Lawson *et al.*, 1986). The Blast-1 molecule has also been proposed as an IgGSF member, with two external units analogous to those of the CD2 chain (one with and one without the disulfide bridge) (Staunton and Thorley-Lawson, 1987). The authors suggest that the Blast-1 sequence is homologous to the α chain of the MHC class II molecule, which also has an N-terminal domain without a disulfide bridge followed by a classical C homology unit. However, the authors indicate homology of the N-terminal region of the Blast-1 chain with those of both the CD4 chain, which is classically V-like, and the MHC class II α chain, sequences that share no apparent similarity themselves. As discussed earlier, if these chains are truly homologous, it is difficult to explain why the CD4 and class II α chains do not appear related. Also, if the crystal structure of the MHC class I molecule proves to be a good model for the structure of the class II molecule, it is apparent that the N-terminal region of the class II α chain does not share any structural homology with Ig domains.

Other molecules that have been proposed as members of the IgGSF include the CD5 molecule (Ly-1) (Huang *et al.*, 1987), a T cell differentiation marker contact site A (cs-A), a cell adhesion molecule from aggregating slime mold (Matsunaga and Mori, 1987), and adenoviral E3 glycoprotein (Chatterjee and Maizel, 1984). We cannot argue that these are not distantly diverged members of the IgGSF. However, none of them meet many of our criteria of membership so we are not convinced it is appropriate to include them as members of the IgGSF. We also believe that the arguments for including the CD2, LFA-3, and Blast-1 molecules are inconclusive as of now.

VI. Evolution of the Immunoglobulin Gene Superfamily

A. Role of the Homology Unit

Given the tremendous diversity within the IgGSF, it might be assumed that selective pressures favoring diversity have driven the evolution of the family. However, we believe that diversity is an inherent feature of the conserved protein and gene structure of the Ig homology unit and that it was the *a priori* diversity itself that shaped the evolution of the complex systems seen today. This diversity is a natural consequence of three basic characteristics of the unit. First, the tertiary structure of the homology unit is such that homology units tend to interact preferentially to form homo- or heterotypic dimers, thereby forming the domains that are the basis of the receptor structures and, at times, receptor/ligand interactions. Combinatorial associations between homology units increase both the evolutionary and somatic diversification potentials. Such interactions may also favor the establishment of new functional associations between existing members of the IgGSF. Second, except for a few conserved residues involved in direct inter- and intrachain interactions, the primary structure of these units can vary dramatically and still result in essentially the same tertiary structure (Lesk and Chothia, 1982). This is easily seen when comparing V gene segments, where even proteins encoded by members of the same gene family may be less than 20% similar. This flexibility is particularly true of the loops connecting the β strands, as seen in the HV regions of Ig and TcR. Such reduced constraints placed by structural requirements on the primary sequence would inherently promote the establishment of variants within populations. Third, in most examples of IgGSF genes for which there is genomic sequence information, the homology units are encoded by discrete exons. This allows, through relatively simple genetic events, the development of new contexts for homology unit expression. For example, most of the known polydomain sequences like N-CAM and FcRγ probably arose independently from internal duplications of unit exons rather than orthologous divergence from a common polydomain precursor. The CEA molecule illustrates that even pairs of exons may duplicate to generate polydomain structures. The number of such independent examples of polydomain sequences indicates the facility of this process. This variation is accommodated by the ½ splicing rule of IgGSF exons, which allows correct splicing between any number of tandem homology unit exons. The ½ splicing rule generally applies as well to the non-Ig-like exons (interestingly, except in the cytoplasmic regions exons) and facilitates the considerable alternative splicing of transcripts seen

throughout the IgGSF. The differential expression of alternately spliced products such as secreted versus membrane Ig and the developmentally expressed variants of N-CAM indicate that alternate RNA splicing provides a further level of functional diversity to the IgGSF. In fact, nearly all members examined to date generate alternate splicing products. Also, the kinase region of PDGFR and CSF-1R and the probable HA binding regions of link protein illustrate the construction of new gene products through the shuffling of exons encoding functionally discrete domains.

B. The Role of Diversity in the History of the IgGSF

While it is likely that many additional members of the IgGSF will be found, we believe general scenarios for the evolution of the IgGSF can be proposed from the current data. We believe that the original homology unit-based molecule was a surface-bound ligand involved in cellular adhesion through binding of like homology units, perhaps nothing more complicated than a cellular glue and probably structurally resembling the Thy-1 (Williams, 1982) or CD3γ chains. The compact structural nature of the homology unit may have contributed to its stability in the extracellular environment. The self-adhesion character of N-CAM and perhaps P_o lends support for a scenario of such homophylic interactions. [It has been suggested that the bisymmetry of the homology unit is an argument for an ancestral "half-domain" structure (Bourgois, 1975). This half-domain would presumably have to form a dimer to function similarly to a current homology unit. Alternatively, its function (or functional capacity) could have been totally unrelated to that of the full domain.] The earliest gene was presumably under selective pressure primarily to maintain the primordial antibody fold and those residues responsible for interchain interactions. Reduced constraint on the primary sequences would promote the establishment of many functionally equivalent alleles. Nonselective emergence of such markers could have been an important preadaptation for several reasons. Differential expression of multiple variants in an individual could result in tissue-specific markers, potentially allowing more complex interactions in morphogenesis to develop. A possible demonstration of a homology unit in a squid brain protein (Williams, 1982) suggests that the IgGSF has probably existed from at least the time of the earliest metazoans and may have an ancient role in morphogenesis. The morphogenic roles of N-CAM, MAG, and P_o may also suggest similar roles for other differentially expressed IgGSF members such as the CD1, T1a, and CEA sequences.

Allelic variants will also represent genotypic markers, and once there is a marker for self, there is conversely an assay for nonself. Recognizing

nonself is important to any organism, such as many invertebrates, that competes with its own kind for space or substrate. Histoincompatibility reactions have been recorded in most multicellular phyla, particularly colonial types (see Hildemann *et al.*, 1980), raising the possibility that MHC-like genes were integral to the development of complex metazoa. Histocompatibility mechanisms could presumably also respond to any "altered-self" markers arising from the interaction of self-histocompatibility molecules and another macromolecular structure. This implies that the immune response is an outgrowth of histocompatibility defenses and could explain the present linkage of MHC and antigen recognition. It has been proposed that the inherent polymorphism was initially incorporated into systems of gametic exclusion to inhibit cross-fertilization of like genotypes and was subsequently harnessed for immune surveillance (Monroy and Rosati, 1979; Burnet, 1971). Interestingly, tunicates have a highly polymorphic gene system responsible for both histocompatibility and gametic exclusion (Scofield *et al.*, 1982). Others suggest that tissue-specific expression of alleles could act as targets for keying developmentally regulated tissue necrosis, a process critical to the development of highly complex metazoa (Williams *et al.*, 1987). Other than the relative order of occurrence, however, none of these scenarios is exclusive of the others; fundamental to each is the *a priori* existence of established polymorphism.

For early members of the IgGSF, tandem duplication of homology units may have been favored initially due to an increase in valency of interaction (i.e., N-CAM). However, subsequent specialization of duplicated regions would also become possible analogous to the divergence of entire duplicated pseudoalleles (Lewis, 1954). A fundamental specialization along these lines appears to have been the establishment of receptor and effector/structural motifs and the subsequent V/C dichotomy. Of course, V and C elements may have become associated through an exon shuffling event long after the divergence of the original elements. Classical V and C homology units have been defined through crystallographic studies of Ig molecules and there has been a tendency to define newly identified members of the IgGSF as either V- or C-like. However, many of the recently described members may not belong to either the V or C group, but rather to a third class of homology unit, H. Though most often described a V-like by primary sequence, the domains of the N-CAM, MAG, PDGFR, CSF-1R FcRγ, α1B, CEA, and CD3 molecules have intercysteine distances even shorter than those of C domains, with the concomitant loss of the extra β-strand/loop structure of V regions as well as possibly another β strand that forms an edge to one of the two β sheets (Fig. 3). The β-strand lengths of the sandwich

faces may tend to be shorter than those of C regions, however, and more nearly like those of V regions. Though more V-like, these sequences also have a distinct collection of conserved residues (Fig. 3). It will be interesting to do a thorough study of the distribution of hydrophobic residues across the faces of the putative β sheets to see if these sequences have a V- or C-like pattern. Such analysis may suggest which face mediates interchain interactions. Given that these sequences tend to share both V and C features, it is possible that they represent more nearly the prototypical homology units and may indicate a more ancient role for the IgGSF in morphogenesis.

C. The Significance of DNA Rearrangement in the Evolution of the IgGSF

The development of regulated rearrangement of V gene segments may have occurred through the opportune insertion and subsequent developmental capture of a transposable element at the 3' portion of a V-like exon (Sakano *et al.*, 1979). With only the two elements, however, this process could have contributed little to receptor diversity. It seems more probable that the initial selective significance of rearrangement was as a mechanism to provide for allelically exclusive regulation of V gene expression. Immune receptor sequences are nearly unique in vertebrates for allelic exclusion. This phenomenon is crucial to the establishment of clonally expressed receptors and, hence, the fine specificity and memory of a complex immune system. Rearrangement provides a mechanism whereby only one V gene is brought near enough to transcriptional promoters in the J-C intron or made available for proper RNA splicing to the C gene. It is also possible that the junctional flexibility of joining may have played initially a more important role in this regulation than in the generation of diversity by further attenuating the possibility of more than one functional rearrangement in one cell (Hunkapiller and Hood, 1989). Once established, allelic exclusion provided an opportunity to express a more specific repertoire of antigen receptors, increasing selective forces on diversity and diversifying mechanisms. Recent characterization of the shark heavy chain locus may support this model (Kobuku *et al.*, 1987). The shark locus heavy chain contains many tandemly repeated V-J repeats, which appear to rearrange only within the linked pairs. This rearrangement preference is essentially the same as for the V_λ loci of mammals. This organization eliminates any combinatorial diversification and reduces the potential for junctional diversity. It is possible, in fact, that it was the acquisition of DNA rearrangement that ultimately drove the divergence of the V and C homology units by allowing the expansion and specialization of the V region repertoire. Furthermore, the development of V-to-J

rearrangement fortuitously established the components that allowed for the evolution of D segments, allowing even further somatic diversification (Hunkapiller and Hood, 1989).

Gene segment rearrangement signals and switch regions also play frequent roles in major recombination events not only within families, but also between families and unrelated sequences on other chromosomes. It has been noted that transposable sequences promote duplication of neighboring sequences and that the rearrangement signal sequences might likewise promote duplicative expansion of V gene segments (Siu *et al.*, 1984). The highly recombinogenic switch sequences of the C_H genes may also lead to expansion of the C gene family as well as the production of novel, hybrid C genes (Hisajima *et al.*, 1983). Translocations involving rearrangement or switch sequences are frequently found in transformed T and B cells (Croce *et al.*, 1984; Tsujimoto *et al.*, 1985). It is unclear whether such events are mediated by the specific enzymes of normal rearrangement or are merely facilitated by the recombinogenic nature of the sequences. The J-like sequences of the CD8 β, CD7, and Ox-2 chains are interesting in this regard (Fig. 9). Unlike most of the single-gene members, the CD8 β (and CD8 α) chain seems to have directly diverged from the Ig light chain V families (Hunkapiller and Hood, 1986). On the other hand, neither the Ox-2 V or C sequence is particularly similar to any of the rearranging gene families. The J-like region of both the CD8 β and Ox-2 genes is encoded within the same exon as the V-like sequence. No germ line information for the CD7 gene is available. Such J-related sequences might be representative of V-like genes that existed before the development of rearrangement, with modern J gene segments the evolutionary descendants of the C-terminal portion of these primordial V genes (Johnson and Williams, 1986). The uniqueness of the Ox-2 (and probably CD7) gene may argue in favor of placing it in this category. The J-like sequence of the CD8 β gene, on the other hand, is nearly identical to some λ J segments, implying a much more recent divergence. Therefore, this J sequence may represent the product of VJ rearrangements in germ cells (Hood *et al.*, 1985). If so, the sequences and perhaps even enzymes involved in somatic variation may occasionally promote rearrangements in germ cells and thus contribute to evolutionary diversification. We feel that this scenario reflects the history of the CD8 β gene and raises interesting questions regarding the role of such events in the evolution of the IgGSF.

D. Implications of the IgGSF

The physical nature and organization of particular genetic information establish the limits and possibilities of its variation. What is striking about the IgGSF is the hierarchical nature of its organization and,

hence, the hierarchical potential of its evolution. The history of the IgGSF reflects both the fundamental diversifying and combinatorial properties of the Ig homology unit. Also, the multigene organization and recombinogenic nature of much of the family establishes particular diversifying potentialities on many of the members. New functional possibilities arise primarily through the duplication of the various informational units involved, e.g., nucleotides, exons, genes, and entire multigene families. Duplication of a multigene family and the attendant cis-acting control mechanisms can in a single event create the genetic basis of a complex new phenotype, suggesting the possibility of rapid evolutionary change. Duplications within the IgGSF not only generate the potential for direct evolutionary change, but also are the seeds of their own subsequent expansion and diversification by the fact that multiple copies of sequences promote further duplication and realignment of gene families. Somatic rearrangement-promoting sequences can apparently also promote such events in the germ line as well. The tendency of even divergent homology units to interact expands even further the potential for rapidly developing new functional relationships between new and existing members. This picture leads to an image of rapid, even saltational evolutionary acquisition of complex new phenotypes that share the regulatory machinery of established systems.

Until recently, members of the IgGSF had been primarily associated with immune recognition or cells of the nervous system. In these systems, complex cell interactions require a surface recognition structure with a high informational potential. The Ig homology unit provides the basic architecture upon which multiple sequence substitutions can be rapidly imposed, thus providing a diverse repertoire of recognition and target elements. The adhesins N-CAM and MAG suggest that other developmentally regulated and cell-specific adhesins may also be dependent on the informational capacity of the homology unit. Interestingly, the number of the identified single-gene IgGSF members that are expressed in both the brain and the immune system suggests the possibility of shared cell surface recognition functions and that related molecules may be involved in some of the intriguing phenomena linking mental states and immune response. It seems likely that the fruitful strategies of somatic diversification employed by well-characterized members of the IgGSF will also be used by other receptor families, perhaps involved in morphogenesis or neuronal development. At present, there is no evidence for this possibility, although candidate systems (e.g., olfaction) have been suggested (Hood *et al.*, 1985). It is thus remarkable to view the entire IgGSF in the context of its evolutionary origins: from a single gene encoding roughly 90 amino acids has arisen a superfamily of

elements with extraordinarily complex interactions both within and outside of the family structure.

REFERENCES

Alt, F. W., and Baltimore, D. (1982). Joining of immunoglobulin heavy chain gene segments. Implications from a chromosome with evidence of 3 D-J-H fusions. *Proc. Natl. Acad. Sci. U.S.A.* **79**, 4118-4122.

Alt, F. W., Yancopoulos, G. D., Blackwell, T. K., Wood, C., Thomas, E., Boss, M., Coffman, R., Rosenberg, N., Tonegawa, S., and Baltimore, D. (1984). Ordered rearrangement of immunoglobulin heavy chain variable region segments. *EMBO J.* 3, 1209-1220.

Amit, A. G., Mariuzza, R. A., Phillips, S. E. V., and Poljack, R. J. (1986). Three-dimensional structure of an antigen-antibody complex at 2.8 Å resolution. *Science* 233, 747-753.

Amzel, L. M., and Poljak, R. J. (1979). Three-dimensional structure of immunoglobulins. *Annu. Rev. Biochem.* **48**, 961-967.

Arden, B., and Klein, J. (1982). Biochemical comparison of major histocompatibility complex molecules from different subspecies of *Mus musculus*: Evidence of transspecific evolution of alleles. *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2342-2346.

Arden, B., Klotz, J. L., Siu, G., and Hood, L. E. (1985). Diversity and structure of genes of the α family of mouse T-cell antigen receptor. *Nature (London)* **316**, 783-787.

Arquint, M., Roder, J., Chia, L.-S., Down, J., Wilkinson, D., Bayley, H., Braun, P., and Dunn, R. (1987). Molecular cloning and primary structure of myelin-associated glycoprotein. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 600-604.

Aruffo, A., and Seed, B. (1987a). Molecular cloning of a CD28 cDNA by a high-efficiency COS cell expression system. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 8573-8577.

Aruffo, A., and Seed, B. (1987b). Molecular cloning of two CD7 (T-cell leukemia antigen) cDNAs by a COS cell expression system. *EMBO J.* **6**, 3313-3316.

Auffray, C., Sikorav, J. L., Ollo, R., and Rougeon, F. (1981). Correlation between D region structure and antigen-binding specificity: Evidences from the comparison of closely related immunoglobulin V_H sequences. *Ann. Immunol.* **132**, 77-88.

Azuma, T., Igras, V., Reilly, E. B., and Eisen, H. N. (1984). Diversity at the variable-joining region boundary of lambda light chains has a pronounced effect on immunoglobulin ligand-binding activity. *Proc. Natl. Acad. Sci. U.S.A.* **81**, 6139-6143.

Baer, R., Chen, K.-C., Smith, S. D., and Rabbits, T. H. (1986). Fusion of an immunoglobulin variable gene and a T-cell receptor constant gene in the chromosome 14 inversion associated with T-cell tumors. *Cell* **43**, 705-714.

Bank, I., DePinho, R. A., Brenner, M. B., Cassimeris, J., Alt, F. W., and Chess, L. (1986). A functional T3 molecule associated with a novel heterodimer on the surface of immature human thymocytes. *Nature (London)* **322**, 179-181.

Barclay, A. N. (1981). Different reticular elements in rat lymphoid tissue identified by localization of Ia, Thy-1 and MRC Ox-2 antigens. *Immunology* **44**, 727-736.

Barclay, A. N., Clark, M. J., and McCaughan, G. W. (1986). Neuronal lymphoid membrane glycoprotein MRC Ox-2 is a member of the immunoglobulin superfamily with a light-chain-like structure. *Biochem. Soc. Symp.* **51**, 149-157.

Barth, R. K., Kim, B. S., Lan, N. C., Hunkapiller, T., Sobieck, N., Winoto, A.,

Gershenfeld, H., Okada, C., Hansburg, D., Weissman, I., and Hood, L. (1985). The murine T-cell receptor employs a limited repertoire of expressed V_β gene segments. *Nature (London)* **316**, 517-523.

Barthels, D., Santoni, M.-J., Willie, W., Ruppert, C., Chaix, J.-C., Hirsch, M.-R., Fontecilla-Camps, J. L., and Goridis, C. (1987). Isolation and nucleotide sequence of mouse NCAM cDNA that codes for a M_r 79,000 polypeptide without a membrane-spanning region. *EMBO J.* **6**, 907-914.

Becker, D. M., Patten, P., Chien, Y.-H., Yokota, T., Eshhar, Z., Giedlin, M., Gascoigne, N. R. J., Goodnow, C., Wolf, R., Arai, K.-I., and Davis, M. M. (1985). Variability and repertoire size of T-cell receptor V_α gene segments. *Nature (London)* **317**, 430-434.

Behlke, M., Spinella, D. G., Chou, H., Sha, W., Hartl, D. L., and Loh, D. Y. (1985). T-cell receptor β chain has limited germline diversity. *Science* **229**, 566-570.

Bernard, A., Boumsell, L., and Hill, C. (1984). *In* "Leucocyte Typing" (A. Bernard, L. Boumsell, J. Dausset, and C. Milstein, eds.), pp. 9-142. Springer-Verlag, Berlin.

Bevan, M. J. (1981). Thymic education. *Immunol. Today* **Nov.**, 216-219.

Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L., and Wiley, D. C. (1987a). Structure of the human class I histocompatibility antigen, HLA-A2. *Nature (London)* **329**, 506-512.

Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L., and Wiley, D. C. (1987b). The foreign antigen binding site and T-cell recognition regions of class I histocompatibility antigens. *Nature (London)* **329**, 512-518.

Bluestone, J. A., Pardall, D. M., Sharron, S. O., and Fowlkes, B. J. (1987). Characterization of murine thymocytes with CD3-associated T-cell receptor structures. *Nature (London)* **326**, 82-84.

Bodmer, W. F., and Bodmer, J. G. (1978). Evolution and function of the HLA system. *Br. Med. Bull.* **34**, 309-316.

Bonnet, F., Périn, J.-P., Lorenzo, F., Jollès, J., and Jollès, P. (1986). An unexpected sequence homology between link proteins of the proteoglycan complex and immunoglobulin-like proteins. *Biochim. Biophys. Acta* **873**, 152-155.

Borst, J., van de Griend, R. J., Van Oostveen, J. W., Ang, S.-L., Melief, C. J., Seidman, J. G., and Bolhurs, R. L. H. (1987). A T-cell receptor γ-CD3 complex found on cloned functional lymphocytes. *Nature (London)* **325**, 683-688.

Bourgois, A. (1975). Evidence for an ancestral immunoglobulin gene coding for half a domain. *Immunochemistry* **12**, 873-876.

Brandtzaeg, P. (1981). Transport models for secretory immunoglobulin A and secretory immunoglobulin M. *Clin. Exp. Immunol.* **44**, 221-232.

Braun, P. E. (1984). Molecular characterization of myelin. *In* "Myelin" (P. Morell, ed.), 2nd Ed., pp. 97-113. Plenum, New York.

Braunstein, N. S., and Germain, R. N. (1987). Allele-specific control of Ia molecule surface expression and conformation: Implications for a general model of Ia structure-function relationships. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 2921-2925.

Bregegere, F. (1983). A directional process of gene conversion is expected to yield dynamic polymorphism associated with stability of alternative alleles in class I histocompatibility antigens gene family. *Biochimie* **65**, 229-237.

Brenner, M. B., McLean, J., Dialynas, D. P., Strominger, J. L., Smith, J. A., Owen, F. L., Seidman, J. G., Ip, S., Rosen, F., and Krangel, M. S. (1986). Identification of a putative second T-cell receptor. *Nature (London)* **322**, 145-149.

Brunet, J.-F., Denizot, F., Luciani, M.-F., Roux-Dosseto, M., Suzan, M., Mattei, M.-G., and Goldstein, P. (1987). A new member of the immunoglobulin superfamily—CTLA-4. *Nature (London)* **328**, 267-270.

Burnet, F. M. (1971). "Self-recognition" in colonial marine forms and flowering plants in relation to the evolution of immunity. *Nature (London)* **232**, 230-235.

Calabi, F., and Milstein, C. (1986). A new family of human MHC-related genes which do not map to chromosome six. *Nature (London)* **323**, 540-543.

Capra, K., and Kehoe, J. M. (1974). Variable region sequences of 5 human immunoglobulin heavy chains of the variable heavy chain. III. Subgroup definitive identification of 4 heavy chain hypervariable regions. *Proc. Natl. Acad. Sci. U.S.A.* **71**, 845-848.

Chatterjee, D., and Maizel, J. V., Jr. (1984). Homology of adenoviral E3 glycoprotein with HLA-DR heavy chain. *Proc. Natl. Acad. Sci. U.S.A.* **81**, 6039-6041.

Chien, Y.-H., Iwashima, M., Kaplan, K. B., Elliot, J. F., and Davis, M. M. (1987a). A new T-cell receptor gene located within the α locus and expressed early in T-cell differentiation. *Nature (London)* **327**, 677-682.

Chien, Y.-H., Iwashima, M., Wettstein, D. A., Kaplan, K. B., Elliot, J. F., Born, W., and Davis, M. M. (1987b). T-cell receptor δ gene rearrangements in early thymocytes. *Nature (London)* **330**, 722-727.

Chothia, C., and Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196**, 901-907.

Chou, P. Y., and Fasman, G. D. (1978). Empirical predictions of protein conformation. *Annu. Rev. Biochem.* **47**, 251-278.

Clark, M. J., Gagnon, J., Williams, A. F., and Barclay, A. N. (1985). MRC OX-2 antigen: A lymphoid/neuronal membrane glycoprotein with a structure like a single immunoglobulin light chain. *EMBO J.* **4**, 113-118.

Clark, S. J., Jefferies, W. A., Barclay, A. N., Gagnon, J., and Williams, A. F. (1987). Peptide and nucleotide sequences of rat CD4 (W3/25) antigen: Evidence for derivation from a structure with four immunoglobulin-related domains. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 1649-1653.

Clarke, S. H., and Rudikoff, S. (1984). Evidence for gene conversion among immunoglobulin heavy chain variable region genes. *J. Exp. Med.* **159**, 773-782.

Clarke, S. H., Huppi, K., Ruezinsky, D., Standt, L., Gerhard, W., and Weigert, M. (1985). Inter- and intraclonal diversity in the antibody response to influenza hemagglutinin. *J. Exp. Med.* **161**, 687-704.

Claverie, J.-M., and Kourilsky, P. (1987). The peptidic self model: A reassessment of the role of the major histocompatibility complex molecules in the restriction of the T-cell response. *Annu. Immunol.* **137**, 425-442.

Cohen, F. E., Novotný, J., Sternberg, M. J. E., Campbell, D. G., and Williams, A. F. (1981). Analysis of structural similarities between brain Thy-1 antigen and immunoglobulin domains. *Biochem. J.* **195**, 31-40.

Colman, P. M., Laver, W. G., Verghese, J. N., Baker, A. T., Tulloch, P. A., Air, G. M., and Webster, R. G. (1987). Three-dimensional structure of a complex of antibody with influenza virus neuraminidase. *Nature (London)* **326**, 358-363.

Concannon, P., Pickering, L. A., Kung, P., and Hood, L. (1986). Diversity and structure of human T-cell receptor β-chain variable region genes. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 6598-6602.

Coussens, L., Van Beveren, C., Smith, D., Chen, E., Mitchell, R. L., Isacke, C. M., Verna, I. M., and Ullrich, A. (1986). Structural alteration of viral homologue of receptor protooncogene *fms* at carboxyl terminus. *Nature (London)* **320**, 277-280.

Cowan, E. P., Jelachich, M. L., Biddison, W. E., and Colligen, J. E. (1987). DNA sequence of HLA-A11: Remarkable homology with HLA-A3 allows identification of residues involved in epitopes recognized by antibodies and T cells. *Immunogenetics* 25, 241-250.

Croce, C. M., Tsujimoto, Y., Eriksen, J., and Nowell, P. (1984). Chromosome translocations and B-cell neoplasia. *Lab. Invest.* **51**, 258–267.

Cunningham, B. A., Wang, J. L., Berggärd, I., and Peterson, P. A. (1973). The complete amino-acid sequence of β_2-microglobulin. *Biochemistry* **12**, 4811–4821.

Cunningham, B. A., Hoffman, S., Rutishauser, U., Hemperly, J. J., and Edelman, G. M. (1983). Molecular topography of the neural cell adhesion molecule N-CAM: Surface orientation and location of sialic acid-rich and binding regions. *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3116–3120.

Cunningham, B. A., Hemperly, J. J., Murray, B. A., Prediger, E. A., Brackenbury, R., and Edelman, G. M. (1987). Neural cell adhesion molecule structure, immunoglobulin-like domains, cell surface modulation and alternative RNA splicing. *Science* **236**, 799–806.

Darsley, M. J., and Rees, A. R. (1985). Nucleotide sequences of 5 anti-lysozyme monoclonal antibodies. *EMBO J.* **4**, 393–398.

Davis, M. M. (1985). Molecular genetics of the T-cell receptor β chain. *Annu. Rev. Immunol.* **3**, 537–560.

Dayhoff, M. O., Barker, W. C., and Hunt, L. T. (1983). Establishing homology in protein sequences. *In*: "Methods in Enzymology" (C. H. W. Hirs and S. N. Timasheff, eds.), Vo. 91, pp. 524–545. Academic Press, New York.

Dembić, Z., Haas, W., Weiss, S., Mc Cubrey, J., Kiefer, H., von Boehmer, H., and Steinmetz, M. (1986). Transfer of specificity by murine α and β T-cell receptor genes. *Nature (London)* **320**, 232–238.

Dembić, Z., Hass, W., Zamoyska, R., Parnes, J., Steinmetz, M., and von Boehmer, H. (1987). Transfection of the CD8 gene enhances T-cell recognition. *Nature (London)* **326**, 510–511.

Desiderio, S. V., Yancopoulos, G. D., Paskind, M., Thomas, E., Bass, M. A., Landau, N., Alt, F. W., and Baltimore, D. (1984). Insertion of N regions into heavy-chain genes is correlated with expression of terminal deoxytransferase in B cells. *Nature (London)* **311**, 752–755.

Diamond, B., and Scharff, M. D. (1984). Somatic mutation of the T15 heavy chain gives rise to an antibody with autoantibody specificity. *Proc. Natl. Acad. Sci. U.S.A.* **81**, 5841–5844.

Dickler, H. B. (1982). Interactions between receptors for antigen and receptors for antibody—A review. *Mol. Immunol.* **19**, 1301–1306.

Dover, G. (1982). Molecular drive: A cohesive mode of species evolution. *Nature (London)* **299**, 111–117.

Early, P., Huang, H., Davis, M., Calane, K., and Hood, L. (1980a). An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: V_H, D and J_H. *Cell* **19**, 981–992.

Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R., and Hood, L. (1980b). Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways. *Cell* **30**, 313–319.

Edelman, G. (1983). Cell adhesion molecules. *Science* **219**, 450–457.

Edelman, G. M. (1986). Cell adhesion molecules in the regulation of animal form and tissue pattern. *Annu. Rev. Cell Biol.* **2**, 81–116.

Edmundson, A. B., Ely, K. R., Abola, E. E., Schiffer, M., and Panagiotopoulos, N. (1975). Rotational allomerism and divergent evolution of domains in immunoglobulin light chains. *Biochemistry* **14**, 3953–3961.

Eisen, H. N. (1986). Why affinity progression of antibodies during immune responses is probably not accompanied by parallel changes in the immunoglobulin-like antigen-specific receptors on T cells. *BioEssays* **4**, 269–272.

Eisen, H. N., and Reilly, E. B. (1985). Lambda chains and genes in inbred mice. *Annu. Rev. Immunol.* **3**, 337-365.

Elliot, J. F., Rock, E. P., Patten, P. A., Davis, M. M., and Chien, Y.-H. (1988). The adult T-cell receptor δ chain is diverse and distinct from that of fetal thymocytes. *Nature (London)* **331**, 627-631.

Fieser, T. M., Tainer, J. A., Geysen, H. M., Houghten, R. A., and Lerner, R. A. (1987). Influence of protein flexibility and peptide conformation on reactivity of monoclonal anti-peptide antibodies with a protein α-helix. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 8568-8572.

Figueroa, F., Golubić, M., Nizetić, D., and Klein, J. (1985). Evolution of the mouse major histocompatibility complex genes borne by *t* chromosomes. *Proc. Natl. Acad. Sci. U.S.A.* **82**, 2819-2823.

Fink, P. J., Matis, L. A., McElligott, D. L., Bookman, M., and Hedrick, S. M. (1986). Correlations between T-cell specificity and the structure of the antigen receptor. *Nature (London)* **321**, 219-226.

Flavell, R. A., Allen, H., Burkly, L. C., Sherman, D. H., Waneck, G. L., and Widera, G. (1986). Molecular biology of the H-2 histocompatibility complex. *Science* **233**, 437-443.

Gabert, J., Langlet, C., Zamoyska, R., Parnes, J. R., Schmitt-Verhulst, A.-M., and Malissen, B. (1987). Reconstitution of MHC-class I specificity by T-cell receptor and Lyt-2 gene transfer. *Cell* **50**, 545-554.

Gallin, W. J., Sorkin, B. C., Edelman, G. M., and Cunningham, B. A. (1987). Sequence analysis of complementary DNA clone encoding the liver cell adhesion molecule L-CAM. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 2808-2812.

Gearhart, P. M., and Bogenhagen, D. F. (1983). Clusters of point mutations are found exclusively around rearranged antibody variable genes. *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3437-3443.

Gearhart, P. M., Johnson, N. D., Douglass, R., and Hood, L. (1981). IgG antibodies to phosphorylcholine exhibit more diversity than their IgM counterparts. *Nature (London)* **291**, 29-34.

Geliebter, J., Zeff, R. A., Melvold, R. Y., and Nathenson, S. (1986). Mitotic recombination in germ cells generated two major histocompatibility complex mutant genes shown to be identical by RNA sequence analysis: K^{bm9} and K^{bm6}. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 3371-3375.

Germain, R. N. (1986). The ins and outs of antigen processing and presentation. *Nature (London)* **22**, 687-689.

Germain, R. N., and Malissen, B. (1986). Analysis of the expression and function of class II major histocompatibility complex-encoded molecules by DNA-mediated gene transfer. *Annu. Rev. Immunol.* **4**, 281-316.

Germain, R. N., and Quill, H. (1986). Unexpected expression of a unique mixed-isotype class II major histocompatibility complex molecule by transfected L cells. *Nature (London)* **320**, 72-75.

Giuillet, J.-G., Lai, M.-Z., Briner, T. J., Smith, J. A., and Gefter, M. L. (1986). Interaction of peptide antigens and class II major histocompatibility. *Nature (London)* **324**, 260-262.

Giusti, A. M., Chien, N. C., Zack, D. J., Shin, S.-U., and Scharff, M. D. (1987). Somatic diversification of S107 from an antiphosphocholine to an anti-DNA autoantibody is due to a single base change in its heavy chain variable region. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 2926-2930.

Gojobori, T., and Nei, M. (1986). Relative contributions of germline gene variation

and somatic mutation to immunoglobulin diversity in the mouse. *Mol. Biol. Evol.* **3**, 156-157.

Gold, D. P., Puck, J. M., Pettey, C. L., Cho, M., Coligan, J., Woody, J. N., and Terhorst, C. (1986). Isolation of cDNA clones encoding the 20K non-glycosylated polypeptide chain of the human T-cell receptor/T3 complex. *Nature (London)* **321**, 431-434.

Gold, D. P., Clevers, H., Alarcon, B., Dunlap, S., Novotný, J., Williams, A. F., and Terhorst, C. (1987). Evolutionary relationship between the T3 chains of the T-cell receptor complex and the immunoglobulin supergene family. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 7649-7653.

Goodhardt, M., Cavelier, P., Akimenko, M. A., Lutfalla, G., Babinet, C., and Rougeon, F. (1987). Rearrangement and expression of rabbit immunoglobulin κ light chain gene in transgenic mice. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 4229-4283.

Gorman, S., Tourvieille, B., and Parnes, J. R. (1987). Structure of the mouse CD4 gene and an unusual transcript in brain. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 7644-7648.

Goverman, J., Minard, K., Shastri, N., Hunkapiller, T., Hansburg, D., Sercarz, E., and Hood, L. (1985). Rearranged β T-cell receptor genes in a helper T-cell clone specific for lysozyme: No correlation between V_β and MHC restriction. *Cell* **40**, 859-867.

Goverman, J., Hunkapiller, T., and Hood, L. (1986). A speculative view of the multicomponent nature of T-cell antigen recognition. *Cell* **45**, 475-484.

Gussow, D., Rein, R. S., Meijer, I., de Hoog, W., Seeman, G. H. A., Hochstenbach, F.M., and Ploegh, H. L. (1987). Isolation, expression and the primary structure of HLA-CW1 and HLA-CW2 genes; evolutionary aspects. *Immunogenetics* **25**, 313-322.

Gustafsson, K., Wiman, K., Emmoth, E., Larhammar, D., Böhme, J., Hyldig-Nielsen, J. J., Ronne, H., Peterson, P. A., and Rask, L. (1984). Mutations and selection in the generation of class II histocompatibility antigen polymorphism. *EMBO J.* **3**, 1655-1661.

Hanau, D., Fabre, M., Schmitt, D. A., Garaud, J.-C., Pauly, G., Tongio, M.-M., Mayer, S., and Cazenave, J.-P. (1987). Human epidermal Langerhans cells cointernalize by receptor-mediated endocytosis "nonclassical" major histocompatibility complex class I molecules (T6 antigens) and class II molecules (HLA-DR antigens). *Proc. Natl. Acad. Sci. U.S.A.* **84**, 2901-2905.

Hannum, C., Freed, J. H., Tarr, G., Kappler, J., and Marrack, P. (1984). Biochemistry and distribution of the T-cell receptor. *Immunol. Rev.* **81**, 161-176.

Hara, T., Fu, S. M., and Hansen, J. A. (1985). Human T-cell activation 2. A new activation pathway used by a major T-cell population via a disulfide-bonded dimer of a 44 Kd polypeptide 9.3 antigen. *J. Exp. Med.* **161**, 1513-1524.

Hayashida, H., and Miyata, T. (1983). Unusual evolutionary conservation and frequent DNA segment exchange in class I genes of the major histocompatibility complex. *Proc. Natl. Acad. Sci. U.S.A.* **80**, 2671-2675.

Hayashida, H., Miyata, T., Honjo, T., Wels, J., Blattner, F., and Yanawaki-Kataoka, Y. (1984). Concerted evolution of the mouse immunoglobulin γ chain genes. *EMBO J.* **3**, 2047-2054.

Hayday, A. C., Diamond, D. J., Tanigawa, G., Heilig, J. S., Folson, V., Saito, H., and Tonegawa, S. (1985a). Unusual organization and diversity of T-cell receptor α-chain genes. *Nature (London)* **316**, 828-832.

Hayday, A. C., Saito, H., Gillies, S. D., Kranz, D. M., Tanigawa, G., Eisen, H. N., and Tonegawa, S. (1985b). The structure organization and somatic rearrangement of T-cell γ genes. *Cell* **40**, 259-269.

Hedrick, S. M., Engel, I., McElligott, D. L., Fink, P. J., Hsu, M.-L., Hansburg, D., and Matis, L. A. (1988). Selection of amino acid sequences in the beta chain of the T cell antigen receptor. *Science* **239**, 1541-1544.

Heidman, O., and Rougeon, F. (1983). Diversity in the rabbit immunoglobulin kappa chain variable regions is amplified by nucleotide deletions and insertions at the V-J junction. *Cell* **34**, 767-777.

Hemperly, J. J., Murray, B. A., Edelman, G. M., and Cunningham, B. A. (1986a). Sequence of a cDNA clone encoding the polysialic acid-rich and cytoplasmic domains of the neural cell adhesion molecule N-CAM. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 3037-3041.

Hemperly, J. J., Edelman, G. M., and Cunningham, B. A. (1986b). Complementary DNA clones of the neural cell adhesion molecule N-CAM lacking a membrane-spanning region consistent with evidence for membrane attachment via a phosphatidylinositol intermediate. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 9822-9826.

Hildemann, W. H., Bigger, C. H., Johnston, J. S., and Jokiel, P. L. (1980). Characteristics of immune memory in invertebrates. *Dev. Immunol.* **10**, 9-14.

Hill, R. L., Delaney, R., Fellow, R. E., Jr., and Lebowitz, H. E. (1966). The evolutionary origins of the immunoglobulins. *Proc. Natl. Acad. Sci. U.S.A.* **56**, 1762-1764.

Hisajima, H., Nishida, Y., Nakai, S., Takahashi, N., Ueda, S., and Honjo, T. (1983). Structure of the human immunoglobulin $C_\epsilon 2$ gene, a truncated pseudogene: Implications for its evolutionary origin. *Proc. Natl. Acad. Sci. U.S.A.* **80**, 2995-2999.

Honjo, T. (1983). Immunoglobulin genes. *Annu. Rev. Immunol.* **1**, 499-528.

Honjo, T., and Habu, S. (1985). Origin of immune diversity, genetic variation and selection. *Annu. Rev. Biochem.* **54**, 803-830.

Hood, L., Gray, W. R., Sanders, B. G., and Dreyer, W. J. (1967). Light chain evolution. *Cold Spring Harbor Symp. Quant. Biol.* **32**, 133-146.

Hood, L., Campbell, J. H., and Elgin, S. C. R. (1975). The organization, expression and evolution of antibody genes and other multigene families. *Annu. Rev. Immunol.* **9**, 305-351.

Hood, L., Steinmetz, M., and Malissen, B. (1983). Genes of the major histocompatibility complex of the mouse. *Annu. Rev. Immunol.* **1**, 529-568.

Hood, L., Kronenberg, M., and Hunkapiller, T. (1985). T-cell antigen receptors and the immunoglobulin supergene family. *Cell* **40**, 225-229.

Howard, J. C. (1987). MHC organization of the rat. Evolutionary considerations. *In* "Evolution and Vertebrate Immunity" (G. Kelsoe and D. H. Schulze, eds.). Univ. of Texas Press, Austin.

Huang, H.-J., Jones, N. H., Strominger, J. L., and Herzenberg, L. A. (1987). Molecular cloning of Ly-1, a membrane glycoprotein of mouse T lymphocytes and a subset of B cells: Molecular homology to its human counterpart Leu-1/T1 (CD5). *Proc. Natl. Acad. Sci. U.S.A.* **84**, 204-208.

Hunkapiller, T., and Hood, L. (1986). Growing immunoglobulin gene superfamily. *Nature (London)* **323**, 15-16.

Hunkapiller, T., and Hood, L. (1989). In preparation.

Ikuta, K., Ogura, T., Shimizu, A., and Honjo, T. (1985). Low frequency of somatic mutation in β-chain variable region genes of human T-cell receptors. *Proc. Natl. Acad. Sci. U.S.A.* **82**, 7701-7705.

Ioannides, C. G., Itoh, K., Fox, F. E., Pahwa, R., Good, R. A., and Platsoucas, C. D. (1987). Identification of a second T-cell antigen receptor in human and mouse by an antipeptide γ-chain-specific monoclonal antibody. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 4244-4248.

Ishioka, N., Takahashi, N., and Putnam, F. W. (1986). Amino acid sequence of human plasma α_1B-glycoprotein: Homology to the immunoglobulin supergene family. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 2363-2367.

Jerne, N. K. (1955). The natural-selection theory of antibody formation. *Proc. Natl. Acad. Sci. U.S.A.* **41**, 849-857.

Johnson, P., and Williams, A. F. (1986). Striking similarities between antigen receptor J pieces and sequence in the second chain of the murine CD8 antigen. *Nature (London)* **323**, 74-76.

Johnson, P., Gagnon, J., Barclay, A. N., and Williams, A. F. (1985). Purification, chain separation and sequence of the MRC OX-8 antigen, a marker of rat cytotoxic T lymphocytes. *EMBO J.* **4**, 2539-2545.

Kaartinen, M., and Mäkelä, O. (1985). Reading of genes in variable frames as a source of antibody diversity. *Immunol. Today* **6**, 324-327.

Kabat, E. A., and Wu, T. T. (1987). *In* "Sequences of Proteins of Immunological Interest" (M. Reid-Miller, H. M. Perry, and K. S. Gottesman, eds.). Public Health Service, NIH, Bethesda, Maryland.

Kelus, A. S., and Weiss, S. (1977). Variant strain of rabbits lacking immunoglobulin κ polypeptide chain. *Nature (London)* **265**, 156-158.

Kikutani, H., Inui, S., Sato, R., Barumian, E. L., Onaki, H., Yamasaki, K., Kaisho, T., Uchibayashi, N., Hardy, R. R., Hirano, T., Tsunasawa, S., Sakiyama, F., Suemura, M., and Kishimoto, T. (1986). Molecular structure of human lymphocyte receptor for immunoglobulin E. *Cell* **47**, 657-666.

Kim, S., Davis, M., Sinn, E., Patten, P., and Hood, L. (1981). Antibody diversity: Somatic hypermutation of rearranged V_H genes. *Cell* **27**, 573-581.

Kimura, M. (1983). "Neutral Theory of Molecular Evolution." Cambridge Univ. Press, London and New York.

Klein, J., and Figueroa, F. (1981). Polymorphism of the mouse H-2 loci. *Immunol. Rev.* **60**, 23-57.

Klein, J., and Figueroa, F. (1986). The evolution of class I MHC genes. *Immunol. Today* **7**, 41-44.

Kleinfield, R., Hardy, R. R., Tarlinton, D., Dangl, J., Herzenberg, L. A., and Weigert, M. (1986). Recombination between an expressed immunoglobulin heavy-chain gene and a germline variable gene segment in a Ly-1-positive B-cell lymphoma. *Nature (London)* **32**, 843-846.

Klinman, N. R. (1976). The acquisition of bone marrow derived cell competence and diversity. *Am. J. Pathol.* **85**, 694-703.

Kodaira, M., Kinashi, T., Umemura, I., Matsuda, F., Noma, T., Ono, Y., and Honjo, T. (1986). Organization and evolution of variable region genes of the human immunoglobulin heavy chain. *J. Mol. Biol.* **190**, 529-541.

Kokubu, F., Hinds, K., Litman, R., Shamblott, M. S., and Litman, G. W. (1987). Extensive families of constant region genes in a phylogenetically primitive vertebrate indicate an additional level of immunoglobulin complexity. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 5868-5872.

Kranz, D. M., Saito, H., Heller, M., Takagaki, Y., Hass, W., Eisen, H. N., and Tonegawa, S. (1985). Limited diversity of the rearranged T-cell γ gene. *Nature (London)* **313**, 752-755.

Krawinkel, U., Zoebelein, G., Brüggemann, M., Radbruch, A., and Rajewsky, K. (1983). Recombination between antibody heavy chain variable-region genes: Evidence for gene conversion. *Proc. Natl. Acad. Sci. U.S.A.* **80**, 4997-5001.

Krissansen, G. W., Owen, M. J., Verbi, W., and Crumpton, M. J. (1986). Primary structure of the T3 γ subunit of the T3/T-cell antigen receptor complex deduced

from cDNA sequences: evolution of the T3 γ and δ subunits. *EMBO J.* 5, 1799-1808.

Kroczek, R. A., Gunter, K. C., Germain, R. N., and Shevach, E. M. (1986). Thy-1 functions as a signal transduction molecule in T lymphocytes and transfected B lymphocytes. *Nature (London)* **322**, 181-184.

Kronenberg, M., Siu, G., Hood, L., and Shastri, N. (1986). The molecular genetics of the T-cell antigen receptor and T-cell antigen recognition. *Annu. Rev. Immunol.* 4, 529-591.

Kühn, L. C., and Kraehenbuhl, J.-P. (1981). The membrane receptor for polymorphic immunoglobulin is structurally related to secretory component: Isolation and characterization of membrane secretory component from rabbit liver and mammary gland. *J. Biol. Chem.* **256**, 12490-12495.

Kunkel, T. A., Gropinathan, K. P., Dube, D. K., Snow, E. T., and Loeb, L. A. (1986). Rearrangements of DNA mediated by terminal transferase. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 1867-1871.

Kuo, C.-L., and Hood, L. (1987). Antigen-major histocompatibility complex-specific activation of murine cells transfected with functionally rearranged T-cell receptor genes. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 7614-7618.

Kurosawa, Y., and Tonegawa, S. (1982). Organization structure and assembly of immunoglobulin heavy chain diversity DNA segments. *J. Exp. Med.* **155**, 201-218.

Lai, C., Brow, M. A., Nave, K.-A., Noronha, A. B., Quarles, R. H., Bloom, F. E., Milner, R. J., and Sutcliffe, J. G. (1987). Two forms of 1B236/myelin-associated glycoprotein, a cell adhesion molecule for post-natal neural development, are produced by alternative splicing. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 4337-4341.

Ledbetter, J. A., Tsu, T. T., and Clark, E. A. (1985). Covalent association between human thymus leukemia-like antigens and CD-8-TP-32 molecules. *J. Immunol.* **134**, 4250-4254.

Lefranc, M.-P., and T. H. Rabbits (1985). Two tandemly organized human genes encoding the T-cell γ constant region sequences show multiple rearrangement in different T-cell lines. *Nature (London)* **316**, 464-466.

Lemke, G., and Axel, R. (1985). Isolation and sequence of a cDNA encoding the major structural protein of peripheral myelin. *Cell* **40**, 501-508.

Lesk, A. M., and Chothia, C. (1982). Evolution of proteins formed by β sheets. *J. Mol. Biol.* **160**, 325-342.

Lew, A. M., Pardell, D. M., Maloy, W. L., Fowlkes, B. J., Kruisbeek, A., Cheng, S.-F., Germain, R. N., Bluestone, J. A., Schwartz, R. H., and Coligan, J. E. (1986). Characterization of a T-cell receptor γ chain expression in a subset of murine thymocytes. *Science* **234**, 1401-1405.

Lewis, E. B. (1954). Pseudoallelism and gene evolution. *Cold Spring Harbor Symp. Quant. Biol.* **16**, 159-174.

Lewis, S., Gifford, A., and Baltimore, D. (1985). DNA elements are asymmetrically joined during the site-specific recombination of kappa immunoglobulin genes. *Science* **228**, 677-685.

Littman, D. R., Thomas, Y., Maddon, P. J., Chess, L., and Axel, R. (1985). The isolation and sequence of the gene encoding T8: A molecule defining functional classes of T lymphocytes. *Cell* **40**, 237-246.

Livant, D., Blatt, C., and Hood, L. (1986). One heavy chain variable region gene segment subfamily in the BALB/c mouse contains 500-1000 or more members. *Cell* **47**, 461-470.

McNicholas, J., Steinmetz, M., Hunkapiller, T., Jones, P., and Hood, L. (1982). DNA sequence of the gene encoding the E_α Ia polypeptide of the BALB/c mouse. *Science* **218**, 1229-1232.

Maddon, P. J., Littman, D. R., Godrey, M., Maddon, D., Chess, L., and Axel, R. (1985). The isolation and nucleotide sequence of a cDNA encoding the T-cell surface protein T4: A new member of the immunoglobulin gene family. *Cell* **42**, 93-104.

Malissen, M., Hunkapiller, T., and Hood, L. (1984). Nucleotide sequence of a light chain of the mouse I-A subregion A_β^d. *Science* **221**, 750-754.

Malissen, B., Shastri, N., Pierres, M., and Hood, L. (1986). Cotransfer of the E_α^d and A_β^d genes into L cells results in the surface expression of a functional mixed-isotype Ia molecule. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 3958-3962.

Martin, L. H., Calabi, F., and Milstein, C. (1986). Isolation of CD1 genes: A new family of MHC-related molecules. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 9154-9158.

Matis, L. A., Cron, R., and Bluestone, J. A. (1987). Major histocompatibility complex-linked specificity of $\gamma\delta$ receptor-bearing T lymphocytes. *Nature (London)* **320**, 262-264.

Matsunaga, T., and Mori, N. (1987). The origin of the immune system. *Scand. J. Immunol.* **25**, 485-495.

Mellor, A. L., Weiss, E. H., Ramachandran, K., and Flavell, R. A. (1983). A potential donor gene for the bm1 gene conversion event in C57BL mouse. *Nature (London)* **306**, 792-795.

Melvold, R. W., and Kohn, H. I. (1975). Histocompatibility gene mutation rates: H-2 and non-H-2. *Mutat. Res.* **27**, 415-417.

Mengle-Gaw, L., and McDevitt, H. O. (1985). Genetics and expression of Ia antigens. *Annu. Rev. Immunol.* **3**, 367-396.

Meuer, S. C., Acuto, O., Hercend, T., Schlossman, S. F., and Reinherz, E. L. (1984). The human T-cell receptor. *Annu. Rev. Immunol.* **2**, 23-50.

Milner, E. C. B., Meek, K. D., Rathbun, G., Tucker, P., and Capra, J. D. (1986). Are antiarsonate antibody N-segments selected at both the protein and the DNA level? *Immunol. Today* **7**, 36-40.

Moingeon, P., Jitsukawa, S., Faure, F., Troglen, F., Triebels, F., Graziani, M., Forestier, F., Bellet, D., Bohuon, C., and Hereend, T. (1987). A γ chain complex forms a functional receptor on cloned human lymphocytes with natural killer-like activity. *Nature (London)* **325**, 723-726.

Monroy, A., and Rosati, F. (1979). The evolution of the cell-cell recognition system. *Nature (London)* **278**, 165-166.

Mostov, K. E., Kraehenbuhl, J.-P., and Blobel, G. (1980). Receptor mediated transcellular transport of immunoglobulin synthesis of secretory component as multiple and larger transmembrane forms. *Proc. Natl. Acad. Sci. U.S.A.* **77**, 7257-7261.

Mostov, K. E., Friedlander, M., and Blobel, G. (1984). The receptor for transepithelial transport of IgA and IgM contains multiple immunoglobulin-like domains. *Nature (London)* **308**, 37-43.

Murre, C., Waldmann, R. A., Morton, C. C., Bongioranni, K. F., Waldman, T. A., Shows, T. B., and Seidman, J. G. (1985). Human γ-chain genes are rearranged in leukemic T cells and map to the short arm of chromosome 7. *Nature (London)* **316**, 549-552.

Nakauchi, H., Shinkai, Y., and Okumura, K. (1987). Molecular cloning of Lyt-3, a membrane glycoprotein marking a subset of mouse T lymphocytes: Molecular homology to immunoglobulin and T-cell receptor variable and joining regions. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 4210-4214.

Neame, P. J., Christner, J. E., and Baker, J. R. (1986). The primary structure of link protein from rat chondrosarcoma proteoglycan aggregate. *J. Biol. Chem.* **261**, 3519-3535.

Nobile-Orazio, E., Hays, A. P., Latov, N., Perman, G., Golier, J., Shy, M. E., and Freddo, L. (1984). Specificity of mouse and human monoclonal antibodies to myelin-associated glycoprotein. *Neurology* **34**, 1336-1342.

Noronha, A. B., Hammer, J. A., Lai, C., Brow, M. A., and Watson, J. B. (1987). Relationship of myelin-associated glycoprotein (MAG) and the brain 1B236 protein. *J. Neurochem.* **48** (suppl.), 533.

Nose, A., Nagafuchi, A., and Takeichi, M. (1987). Isolation of placental cadherin cDNA: Identification of a novel gene family of cell-cell adhesion molecules. *EMBO J.* **6**, 3655-3661.

Novotný, J., Bruccoleri, R., and Karplus, M. (1984). An analysis of incorrectly folded protein models: Implications for structure predictions. *J. Mol. Biol.* **177**, 787-818.

Novotný, J., Tonegawa, S., Saito, H., Kranz, D., and Eisen, H. (1986). Secondary, tertiary, and quaternary structure of T-cell-specific immunoglobulin-like polypeptide chains. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 742-746.

O'Brien, S. J., Roelke, M. E., Marker, L., Newman, A., Winkler, C. A., Meltzer, D., Colly, L., Evermann, J. F., Bush, M., and Wildt, D. E. (1985). Genetic basis for species vulnerability in the cheetah. *Science* **227**, 1428-1433.

O'Brien, S. J., Wildt, D. E., Bush, M., Caro, T. M., FitzGibbon, C., Aggundey, I., and Leakey, R. E. (1987). East African cheetahs: Evidence for two population bottlenecks? *Proc. Natl. Acad. Sci. U.S.A.* **84**, 508-511.

Ohashi, P. S., Mak, T. W., Van den Elsen, P., Yanagi, Y., Yoshikai, Y., Calman, A. F., Terhorst, C., Stobo, J. D., and Weiss, A. (1985). Reconstitution of an active surface T3/T-cell antigen receptor by DNA transfer. *Nature (London)* **316**, 606-609.

Ohnishi, K. (1984). Domain structures and molecular evolution of class I and class II major histocompatibility gene complex. MHC products deduced from amino acid and nucleotide sequence homologies. *Origins Life* **14**, 707-716.

Ohta, T. (1982). Allelic and nonallelic homology of a supergene family. *Proc. Natl. Acad. Sci. U.S.A.* **79**, 3251-3254.

Palker, T. J., Searle, R. M., Hensley, L. L., Ho, W., and Haynes, B. F. (1985). *In* "Leukocyte Typing II" (E. L. Reinherz, B. F. Haynes, L. M. Nadler, and I. D. Berstein, eds.), pp. 303-313. Springer-Verlag, New York.

Pardoll, D. M., Fowlkes, B. J., Bluestone, J. A., Kruisbeek, A., Maley, W. L., Colian, J. E., and Schwartz, R. H. (1987). Differential expression of two distant T-cell receptors during thymocyte development. *Nature (London)* **326**, 79-81.

Parnes, J. R., and Hunkapiller, T. (1987). L3T4 and the immunoglobulin gene superfamily: New relationships between the immune system and the nervous system. *Immunol. Rev.* **100**, 109-127.

Parnes, J. R., and Seidman, J. G. (1982). Structure of wild type and mutant mouse β_2-microglobulin genes. *Cell* **29**, 661-670.

Patten, P., Yokota, T., Rothbard, J., Chien, Y., Arai, K., and Davis, M. (1984). Structure, expression and divergence of T-cell receptor β chain variable regions. *Nature (London)* **312**, 40-46.

Paxton, R., Mooser, G., Pande, H., Lee, T. D., and Shively, J. E. (1987). Sequence analysis of carcinoembryonic antigen: Identification of glycosylation sites and homology with the immunoglobulin supergene family. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 920-924.

Pease, L. R., Schulze, D. H., Pfaffenbach, G. M., and Nathenson, S. G. (1983). Spontaneous H-2 mutants provide evidence that a copy mechanism analogous to gene conversion generates polymorphism in the major histocompatibility complex. *Proc. Natl. Acad. Sci. U.S.A.* **80**, 242-246.

Peault, B., Chen, C.-L. H., Cooper, M. D., Barbu, M., Lipinski, M., and Le Douarin, N. M. (1987). Phylogenetically conserved antigen on nerve cells and lymphocytes resembles myelin-associated glycoprotein. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 814–818.

Poggi, A., Botlino, C., Zoochi, M. R., Pantaleo, G., Ciccone, E., Mingari, C., Moretta, L., and Moretta, A. (1987). CD3-positive WT3]-negative peripheral T lymphocytes lack T44 CD28, a surface molecule involved in activation of T cells bearing the α-β heterodimer. *Eur. J. Immunol.* **17**, 1065–1068.

Quarles, R. H. (1984). Myelin-associated glycoprotein in development and disease. *Dev. Neurosci.* **6**, 285–303.

Quartermous, T., Strauss, W., Murre, C., Dialynas, D. P., Strominger, J. L., and Seidman, J. G. (1986). Human T-cell γ genes contain N segments and have marked junctional variability. *Nature (London)* **322**, 184–187.

Ravetch, J. V., Luster, A. D., Weinshank, R., Kochan, J., Pavlovec, A., Portnoy, D. A., Hulmes, J., Pan, Y.-C. E., and Unkeless, J. C. (1986). Structural heterogeneity and functional domains of murine immunoglobulin G Fc receptor. *Science* **234**, 718–725.

Reinherz, E. L. (1985). A molecular basis for thymic selection: Regulation of T11 induced thymocyte expansion by the T3-T1 antigen/MHC receptor pathway. *Immunol. Today* **3**, 75–76.

Reynaud, C.-A., Angques, V., Dahan, A., and Weill, J. C. (1985). A single rearrangement event generates most of the chicken immunoglobulin light chain diversity. *Cell* **40**, 283–292.

Richardson, J. S., Richardson, D. C., Thomas, K. A., Silverton, E. W., and Davies, D. R. (1976). Similarity of 3-dimensional structure between the immunoglobulin domain and the copper-zinc super oxide dimatase subunit. *J. Mol. Biol.* **102**, 221–235.

Ringwald, M., Schuh, R., Vestweber, D., Eistetter, H., Lottspeich, F., Engel, J., Dölz, R., Jähnig, F., Epplen, J., Mayer, S., Müller, C., and Kemler, R. (1987). The structure of cell adhesion molecule uvomorulin. Insights into the molecular mechanism of Ca^{2+}-dependent cell adhesion. *EMBO J.* **6**, 3647–3653.

Riopelle, R. J., McGarry, R. C., and Roder, J. C. (1986). Adhesion properties of a neuronal epitope recognized by the monoclonal antibody HNK-1. *Brain Res.* **367**, 20–25.

Rogers, J. (1985). Mouse histocompatibility-related genes are not conserved in other mammals. *EMBO J.* **4**, 749–754.

Rudikoff, S., and Pumphrey, J. A. (1986). Functional antibody lacking a variable-region disulfide bridge. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 7875–7878.

Rudikoff, S., Pawlita, M., Pumphrey, J., and Heller, M. (1984). Somatic diversification of immunoglobulins. *Proc. Natl. Acad. Sci. U.S.A.* **81**, 2162–2166.

Sablitzky, F., Wildner, G., and Rajewsky, K. (1985). Somatic mutation and clonal expansion of B cells in an antigen-driven immune response. *EMBO J.* **4**, 345–350.

Saito, H., Kranz, D. M., Takagaki, Y., Hayday, A. C., Eisen, H. N., and Tonegawa, S. (1984). Complete primary structure of a heterodimeric T-cell receptor deduced from cDNA sequences. *Nature (London)* **309**, 757–762.

Sakano, H., Huppi, K., Heinrich, G., and Tonegawa, S. (1979). Sequences of the somatic recombination sites of immunoglobulin light chain genes. *Nature (London)* **280**, 288–294.

Sakano, H., Maki, R., Kurosawa, Y., Roeder, W., and Tonegawa, S. (1980). 2 types of somatic recombination are necessary for the generation of complete immunoglobulin heavy chain genes. *Nature (London)* **286**, 676–683.

Schultze, H. E., Heide, H., and Haupt, H. (1963). Isolation of an easily percipitable α_1-glycoprotein of human serum. *Nature (London)* **200**, 1103–1103.

Schwartz, R. H. (1984). The role of gene products of the major histocompatibility complex in T-cell activation and cellular interactions. *In* "Fundamental

Immunology" (W. E. Paul, ed.), pp. 379-438. Raven Press, New York.

Scofield, V. L., Schlumpberger, J. M., West, L. A., and Weissman, I. L. (1982). Protochordate allorecognition is controlled by a MHC-like gene system. *Nature (London)* **295**, 499-502.

Seed, B. (1987). An LFA-3 complementary DNA encodes a phospholipid-linked membrane protein homologous to its receptor CD2. *Nature (London)* **329**, 840-842.

Sewell, W. A., Brown, M. H., Dunne, J., Owen, M. J., and Crumpton, M. J. (1986). Molecular cloning of the human T-lymphocyte surface CD2 (T11) antigen. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 8718-8722.

Sheriff, S., Silverton, E. W., Padlan, E. A., Cohen, G. H., Smith-Gill, S. J., Finzell, B. C., and Davies, D. R. (1987). Three-dimensional structure of an antibody-antigen complex. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 8075-8079.

Shimizu, A., Takahashi, N., Yaota, Y., and Honjo, T. (1982). Organization of the constant region gene family of the mouse immunoglobulin heavy chain. *Cell* **28**, 499-506.

Shively, J. E., and Beatty, J. D. (1985). CEA-related antigens: Molecular biology and clinical significance. *CRC Crit. Rev. Oncol. Hematol.* **2**, 355-399.

Siu, G., Kronenberg, M., Strauss, E., Haars, R., Mak, T. W., and Hood, L. (1984). The structure, rearrangement and expression of D_β gene segments of the murine T-cell antigen receptor. *Nature (London)* **311**, 344-350.

Snow, P. M., Van de Rijn, M., and Terhorst, C. (1985). Association between the human thymic differentiation antigens T-6 and T-8. *Eur. J. Immunol.* **15**, 529-532.

Soloski, M. J., Vernachio, J., Emhorn, G., and Lattimore, A. (1986). Qa gene expression, biosynthesis and secretion of Qa-2 molecules in activated T cells. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 2949-2953.

Springer, T. A., and Unkeless, J. C. (1984). Analysis of macrophage differentiation and function with monoclonal antibodies. *Contemp. Top. Immunobiol.* **13**, 1-31.

Staunton, D. E., and Thorley-Lawson, D. A. (1987). Molecular cloning of the lymphocyte activation marker Blast-1. *EMBO J.* **6**, 3695-3701.

Stephan, D., Sun, H., Fischer-Lindahl, K., Meyer, E., Hämmerling, G., Hood, L., and Steinmetz, M. (1986). Organization and evolution of D region class I genes in the mouse major histocompatibility complex. *J. Exp. Med.* **163**, 1227-1244.

Sternberger, N. H., Quarles, R. H., Itoyama, Y., and Webster, H. deF. (1979). Myelin-associated glycoprotein demonstrated immunocytochemically in myelin and myelin-forming cells of developing rat. *Proc. Natl. Acad. Sci. U.S.A.* **76**, 1510-1514.

Streilein, J. W., and Duncan, W. R. (1983). On the anomalous nature of the major histocompatibility complex in Syrian hamsters, Hm-1. *Transplant. Proc.* **15**, 1540-1545.

Strominger, J. L., Orr, H. T., Parham, P., Ploegh, H. L., Mann, D. L., Bilofsky, H., Saroff, H. A., Wu, T. T., and Kabat, E. A. (1980). An evaluation of the significance of amino-acid sequence homologies in human histocompatibility antigens HLA-A and HLA-B with immunoglobulins and other proteins using relatively short sequences. *Scand. J. Immunol.* **11**, 573-592.

Sukhatme, V. P., Sizer, K. C., Vollmer, A. C., Hunkapiller, T., and Parnes, J. R. (1985). The T-cell differentiation antigen Leu-2/T8 is homologous to immunoglobulin and T-cell receptor variable regions. *Cell* **40**, 591-597.

Sun, Y. H., Goodenow, R. S., and Hood, L. (1985). Molecular basis of the dnl mutation in the major histocompatibility complex of the mouse: A D/L hybrid gene. *J. Exp. Med.* **162**, 1588-1602.

Takai, Y., Reed, M. L., Burgkoff, S. J., and Hermann, S. H. (1987). Direct evidence for a receptor-ligand interaction between the T-cell surface antigen CD2 and lymphocyte-function-associated antigen 3. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 6864-6868.

Terhorst, C., Georgopoulos, K., Gold, D., Oettgen, H., Pettey, C., Ucker, D., and Van den Elsen, P. (1986). *In* "Regulation of Immune Gene Expression" (M. Feldman and A. McMichael, eds.). Humana Press, New York.

Thompson, J. A., Pande, H., Paxton, R. J., Shively, L., Padma, A., Simmer, R. L., Todd, C. W., Riggs, A. D., and Shively, J. E. (1987). Molecular cloning of a gene belonging to the carcinoembryonic antigen gene family and discussion of a domain model. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 2965-2969.

Thorley-Lawson, D. A., Schooley, R. T., Bhan, A. K., and Nadler, L. M. (1982). Epstein Barr virus super induces a new human B cell differentiation antigen Blast-1 expressed on transformed lymphoblasts. *Cell* **30**, 415-425.

Thorley-Lawson, D. A., Swendemon, S., and Edson, L. M. (1986). Biochemical analysis suggests distinct functional roles for the Blast-1 and Blast-2 antigens. *J. Immunol.* **136**, 1745-1751.

Tonegawa, S. (1983). Somatic generation of antibody diversity. *Nature (London)* **302**, 575-581.

Tourvieille, B., Gorman, S. D., Field, E. H., Hunkapiller, T., and Parnes, J. R. (1986). Isolation and sequence of an L3T4 cDNA clone: Expression in T cells and brain. *Science* **234**, 610-614.

Townsend, A. (1985). Molecules at work on the T-cell surface. *Immunol. Today* **6**, 68-70.

Tse, A. G. D., Barclay, A. N., Watts, A., and Williams, A. F. (1985). A glycophospholipid tail at the carboxyl-terminus of the Thy-1 glycoprotein of neurons and thymocytes. *Science* **230**, 1003-1008.

Tsujimoto, Y., Gorham, J., Cossman, J., Jaffe, E., and Croce, C. M. (1985). The 14.18 chromosome translocations involved in B cell neoplasm result from mistakes in VDJ joining. *Science* **229**, 1390-1393.

Van den Elsen, P., Shepley, B.-A., Borst, J., Coligan, J. E., Markham, A. F., Orkin, S., and Terhorst, C. (1984). Isolation of cDNA clones encoding the 20K T3 glycoprotein of human T-cell receptor complex. *Nature (London)* **312**, 413-418.

Vice, J. L., Gilman-Sachs, A., Hunt, W. L., and Dray, S. (1970). Allotype suppression in a^2a^2 homozygous rabbits fostered *in uteri* of a2-immunized a^1a^1 homozygous mothers and injected at birth with anti-a2 antiserum. *J. Immunol.* **104**, 550-554.

Wabl, M., Burrows, P. D., von Gabain, A., and Steinberg, C. (1985). Hypermutation at the immunoglobulin heavy chain locus in a pre-B-cell line. *Proc. Natl. Acad. Sci. U.S.A.* **82**, 479-482.

Wake, C. T. (1986). Molecular biology of the HLA class I and class II genes. *Mol. Biol. Med.* **3**, 1-11.

Wallner, B. P., Frey, A. Z., Tizard, R., Mattaliano, R. J., Hession, C., Sanders, M. E., Dustin, M. L., and Springer, T. A. (1987). Primary structure of lymphocyte-function-associated antigen 3 LFA-3, the ligand of the T lymphocyte CD2 glycoprotein. *J. Exp. Med.* **166**, 923-932.

Webb, M., and Barclay, A. N. (1984). Localization of the MRC OX-2 glycoprotein on the surfaces of neurons. *J. Neurochem.* **43**, 1061-1067.

Weill, J. C. (1986). Generation of diversity in the avian bursa. *Prog. Immunol.* **6**, 20-25.

Weiss, E. H., Mellor, A., Golden, L., Fahrner, K., Simpson, E., Hurst, J., and Flavell, R. A. (1983). The structure of a mutant H-2 gene suggests that the generation of polymorphism in H-2 genes may occur by gene conversion-like events. *Nature (London)* **301**, 671-674.

Westhof, E., Altschuh, D., Moras, D., Bloomer, C., Mondragon, A., Klug, A., and von Regenmortel, M. H. V. (1984). Correlation between segmental mobility and the location of antigenic determinants in proteins. *Nature (London)* **311**, 123-126.

Williams, A. F. (1982). Surface molecules and cell interactions. *J. Theor. Biol.* **98**, 221-231.

Williams, A. F. (1984). The immunoglobulin superfamily takes shape. *Nature (London)* **308**, 12-13.

Williams, A. F. (1987). A year in the life of the immunoglobulin superfamily. *Immunol. Today* **8**, 298-303.

Williams, A. F., Barclay, A. N., Clark, S. J., Paterson, D. J., and Willis, A. C. (1987). Similarities in sequences and cellular expression between rat CD2 and CD4 antigens. *J. Exp. Med.* **165**, 368-380.

Wilson, R. K., Lai, E., Concannon, P., Barth, R. K., and Hood, L. E. (1988). Structure, organization and polymorphism of murine and human T-cell receptor α and β chain gene families. *Immunol. Rev.* **101**, 149-172.

Winoto, A., Mjoleness, S., and Hood, L. (1985). Genomic organization of the genes encoding mouse T-cell receptor α-chain. *Nature (London)* **316**, 832-836.

Winoto, A., Urban, J. L., Lan, N. C., Goverman, J., Hood, L., and Hansburg, D. (1986). Predominant use of a Vα gene segment in mouse T-cell receptors for cytochrome c. *Nature (London)* **324**, 679-682.

Wolff, K., and Stingl, G. (1983). The Langerhans cell. *J. Invest. Dermatol.* **80**, 17S-21S.

Wu, T. T., and Kabat, E. A. (1970). An analysis of the sequences of the variable regions of Bence-Jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med.* **132**, 211-250.

Xu, C.-S., Tung, E., Wang, I. Y., and Wang, A. C. (1987). Identification and characterization of Thy-1 homologous from bovine thymocytes. *Mol. Immunol.* **24**, 791-796.

Yagüe, J., White, J., Coleclough, C., Kappler, J., Palmer, E., and Marrack, P. (1985). The T-cell receptor: The α and β chains define idiotype, and antigen and MHC specificity. *Cell* **42**, 81-87.

Yancopoulos, G., Blackwell, K., Suh, H., Hood, L., and Alt, F. (1986). Introduced T-cell receptor variable region gene segments recombine in pre-B cells. Evidence that B and T cells use a common recombinase. *Cell* **44**, 251-260.

Yarden, Y., Escobedo, J. A., Kuang, W.-J., Yang-Feng, T. L., Daniel, T. O., Tremble, P. M., Chen, E. Y., Ando, M. E., Harkins, R. N., Francke, U., Fried, V. A., Ulrich, A., and Williams, L. T. (1986). Structure of the receptor for platelet-derived growth factor helps define a family of closely-related growth factor receptors. *Nature (London)* **323**, 226-232.

Yoshikai, Y., Anatoniou, D., Clark, S. P., Yanagi, Y., Sangster, R., Van den Elsen, P., Terhorst, C., and Mak, T. W. (1984). Sequence and expression of transcripts of the human T-cell receptor β-chain genes. *Nature (London)* **312**, 521-524.

Yoshikai, Y., Clark, S. P., Taylor, S., Sohn, U., Wilson, B., Minden, M., and Mak, T. W. (1985). Organization and sequences of the variable, joining and constant region genes of the human T-cell receptor α chain. *Nature (London)* **316**, 837-840.

Zamoyska, R., Vollmer, A. C., Sizer, K. C., Liaw, C. W., and Parnes, J. R. (1985). Two Lyt-2 polypeptides arise from a single gene by alternative splicing patterns of mRNA. *Cell* **43**, 153-163.

Ziegler, A., and Milstein, C. (1979). A small polypeptide different from β-2 microglobulin associated with a human cell surface antigen. *Nature (London)* **279**, 243-244.

Zimmerman, W., Ortlieb, B., Friedrich, R., and von Kleist, S. (1987). Isolation and characterization of cDNA clones encoding the human carcinogenic antigen reveal a highly conserved repeating structure. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 2960-2964.

Cell, Vol. 45, 475–484, May 23, 1986.

Review

A Speculative View of the Multicomponent Nature of T Cell Antigen Recognition

Joan Goverman, Tim Hunkapiller, and Leroy Hood
Division of Biology
California Institute of Technology
Pasadena, California 91125

Introduction

Recent data on the molecular biology of T cells have given us a detailed view of the genes and proteins involved in antigen recognition by T cells. T and B cells are responsible for specific antigen recognition in the vertebrate immune response. B cells recognize free antigen through cell surface-bound immunoglobulin (Ig) with no other requirement except antigen/receptor complementarity. T cells recognize antigen only when it is presented on the cell surface in the context of class I or class II molecules encoded by the major histocompatibility complex (MHC). This interaction is said to be self-MHC restricted in that only antigen presented on cells expressing self-MHC alleles can be recognized. Three functional classes of T cells have been defined: T helper (T_H) cells, which stimulate immune responses; T suppressor (T_S) cells, which diminish these responses; and T cytotoxic (T_C) cells, which are involved in the direct killing of cells expressing antigenic molecules. T_H cells usually recognize antigen associated with class II MHC molecules expressed selectively on antigen presenting cells (APC) such as B cells and macrophages. T_C cells generally respond to antigens of cellular origin, such as viral or tumor antigens in association with the class I molecules present on most cells. T_S cells are less well characterized and will not be further considered in this review. Various models have been proposed to provide a conceptual framework for the unique requirements of T cell antigen recognition. We present here a model that is in many ways a synthesis of some of these earlier views, but expands upon them in an attempt to rationalize the known cellular constraints with those imposed by the new molecular data. This paper reflects our point of view and is not meant to be a comprehensive review of the literature.

Biological Constraints

There are five important phenomena that must be addressed by any model of antigen recognition by T cells. First, antigen recognition appears to require simultaneous recognition of self-MHC molecules (Kindred and Shreffler, 1972; Katz et al., 1973a; Doherty and Zinkernagel, 1975). Second, chimera studies with T cell precursor populations (bone marrow) and thymus grafts suggest that T cell populations are "educated" as to what constitutes self during maturation in the thymus (for review, see Longo et al., 1981). It has been proposed that this results from a process of selection for T cells that recognize only the collection of self-MHC alleles of thymic origin (Fink and Bevan, 1978; Waldman et al., 1978; Zinkernagel et al., 1978). Third, in generating a T cell repertoire that recognizes foreign antigen in association with self-MHC, receptors specific for self-MHC molecules alone must be eliminated. However, a large fraction of T cells recognize allo-MHC molecules alone (Ford et al., 1975; Fischer-Lindahl and Wilson, 1977). Since vertebrate immune systems never see allogeneic cells under normal circumstances, the frequency and intensity of allo-responses presumably reflect some fundamental property of the T cell receptor. Fourth, animals that generate otherwise normal immune responses can be nonresponders to particular antigens. Genes controlling these immune response (Ir) effects encode the class I and class II MHC molecules (for review, see Schwartz, 1984). Fifth, the class of restricting MHC molecules generally correlates with T cell effector function, class I with T_C cells and class II with T_H cells (Katz et al., 1973b; Zinkernagel and Doherty, 1975; Kappler and Marrack, 1976).

Molecular Constraints

Each T cell expresses a clone-specific receptor, a heterodimer of α and β glycoproteins (Allison et al., 1982; Haskins et al., 1983; Meuer et al., 1983). In humans and mice the α/β complex noncovalently associates with T3, a multi-protein cell-surface complex (Meuer et al., 1983; van den Elsen et al., 1985). Expression of both α and β chains is required for MHC and antigen specificity (Yagüe et al., 1985). Each chain, like the Ig light (L) and heavy (H) chains, has a variable (V) and constant (C) region (Acuto et al., 1983; Kappler et al., 1983; McIntyre and Allison, 1983). These regions are similar to Ig homology units, and hence α and β genes belong to the Ig gene superfamily (for review, see Hood et al., 1985). The V_β genes are encoded by multiple V_β, diversity (D_β), and junctional (J_β) gene segments that rearrange during differentiation to form a contiguous gene (for reviews, see Davis, 1985; Kronenberg et al., 1986). V_α genes rearrange V_α and J_α gene segments, but no D_α gene segments have been demonstrated (Chien et al., 1984; Saito et al., 1984a; Sim et al., 1985; Arden et al., 1985; Hayday et al., 1985a; Winoto et al., 1985). Statistical analysis indicates that in mice there are about 100 V_α gene segments (calculation not shown) and about 30 V_β gene segments (Barth et al., 1985; Behlke et al., 1985). At least 30, and probably greater than 50, J_α gene segments are dispersed over 60 kb preceding a single C_α gene (Hayday et al., 1985a; Winoto et al., 1985). There are two linked C_β genes, each preceded by one D_β gene segment and a cluster of six J_β gene segments (Gascoigne et al., 1984; Malissen et al., 1984). The two C_β sequences appear to be functionally equivalent (Kronenberg et al., 1985). It appears that essentially any $V_\beta D_\beta J_\beta$ or $V_\alpha J_\alpha$ combination is possible, except that $D_{\beta 2}$ probably only rearranges to the $J_{\beta 2}$ cluster. Apparently, only one functional α and one functional β chain is expressed per cell, a phenomenon analogous to

Table 1. Combinatorial and Somatic Diversification of Mouse V Genes

	alpha	beta	Heavy	kappa
V gene segments (subfamilies)	100 (11)	30 (14)	500 (7)	200 (5)
D gene segments	–	2	15	–
J gene segments	50	5+6	4	4
D's in 3 reading frames	–	+++	+	–
N region sequence	+	++	++	–
Junctional diversity	+++	+++	+++	+
Somatic hypermutation	–	–	+	+
Alternate joining order	–	+?	–	–
Combinatorial joining	VxJ 100x50	VxDxJ (30x3x12) + (30x3x6)	VxDxJ 500x15x4	VxJ 200x4
Total	5000	x 1620	3×10^4	x 800
Combinatorial association		8.1×10^6		2.4×10^7

The number of V gene segments represents calculated estimates of germ line segments modified here to accommodate a 25%–50% incidence of pseudogenes. The number of + 's indicates the relative contribution of the indicated mechanism to diversity between the different V gene families. Members of subfamilies are assumed to be about 80% similar at the DNA level. The number of subfamilies listed are those known from the estimated V gene segment sample size and are included to indicate the degree of variability between the members of any one family. The contribution of somatic hypermutation to the B cell repertoire may function more to "fine-tune" secondary responses to particular antigen responses rather than to extend the basic repertoire (Gearhart et al., 1981; Herron and Voss, 1983).

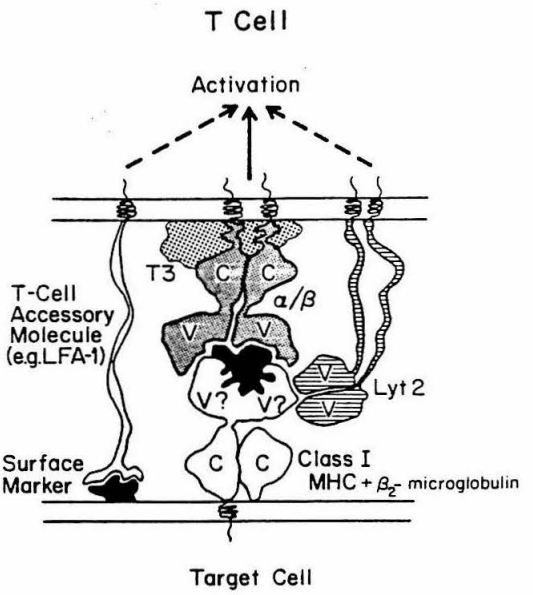


Figure 1. Model of Interaction-antigen Recognition by T Cells

Different molecules are indicated by different shadings. Immunoglobulin-like regions are indicated by V and C. The interaction-antigen is formed by the complex of the black antigen with the membrane distal domain of the MHC class I molecule. The T cell α/β receptor is complementary to the complex. Lyt2 recognizes a nonpolymorphic structure on the distal domain of the MHC molecule and may direct the α/β receptor to antigen which is presented by that same MHC molecule. LFA-1 is included as an example of T cell-specific molecules potentially involved in T cell activation. Antigen recognition involving MHC class II molecules is presumably similar to that shown for class I, but would involve L3T4 instead of Lyt2.

the allelic exclusion of Ig genes in B cells (Luzzati et al., 1973; Goverman et al., 1985; Kronenberg et al., 1985; Malissen et al., 1985). Both α and β genes share with Ig genes several somatic variation processes directly linked to rearrangement (for reviews, see Tonegawa, 1983; Kronenberg et al., 1986). As indicated in Table 1, somatic hypermutation has been seen only in B cells, but may be involved more in affinity maturation than in generating new B cell specificities. We believe that consideration of combinatorial possibilities, gene-segment germ line diversity, and somatic variation indicates that T cells are capable of expressing at least as diverse a receptor population as B cells (Table 1). Another T cell-specific gene, γ, is constructed by rearranging gene segments similar to α and β genes (Saito et al., 1984b; Hayday et al., 1985b). No γ protein has been demonstrated, however, and the potential diversity of V_γ sequences appears to be limited (Kranz et al., 1985). This suggests to us that γ does not play an important role in antigen recognition, although other views have been proposed (Pernis and Axel, 1985; Raulet et al., 1985).

T cells also express accessory molecules that define distinct subsets of the mature T cell population. L3T4 and Lyt2 are two such molecules in mice whose expression correlates with T_H and T_C cell lineages, respectively (Swain, 1981; Meuer et al., 1982; Marrack et al., 1983; Spits et al., 1985). The Lyt2 gene and those of the human equivalents of Lyt2 and L3T4, T8 and T4, are single-copy, nonpolymorphic genes. Each contains a V-like homology unit and hence belongs to the Ig gene superfamily and may form V domain-like receptors for nonpolymorphic determinants (Littman et al., 1985; Maddon et al., 1985; Sukhatme et al., 1985; Zamoyska et al., 1985).

Model

Our view of antigen recognition by T cells has five tenets. First, the α/β receptor recognizes novel determinants of a complex formed by interaction of an antigen and MHC molecule. The receptor does not distinguish the antigen or MHC source of the structures forming these determinants. Consequently, T cell receptors do not identify self-MHC-specific determinants, and individual T cells are not a priori limited to recognizing antigen only presented by self-MHC molecules. Second, T cells are not specifically selected during development to recognize self-MHC determinants, either preferentially or exclusively. Thymic-MHC restriction is a limited bias imposed on a T cell population both by interaction-antigen-selected clonal triggering and non-allele-specific, low-affinity interaction with MHC molecules alone. Exposure of T cells to the same antigen and/or MHC at different points in their development may play a role in both the establishment of self-tolerance

and the expansion of a peripheral T cell population. Third, the class of an antigen-presenting MHC molecule is functionally defined through T cell recognition of MHC sequences different from those involved in antigen presentation. These sequences are recognized by accessory molecules such as Lyt2 and L3T4, and play a role in determining T cell effector function. Fourth, binding of an interaction antigen by the α/β complex is required but not sufficient for T cell activation. Other factors or signals generated by the cellular presentation of antigen are generally required; thus, binding of free antigen is possible but will not result in T cell activation. Fifth, all structures involved in specific antigen or MHC recognition belong to the Ig gene superfamily and consequently employ similar recognition strategies.

The basis of this model is that the α/β receptor recognizes an antigen/MHC complex in essentially the same way that the Ig H/L dimer recognizes antigen alone. In order for T cell responses to be limited to cellularly presented antigens, some cell-surface structure must be recognized independently of antigen. MHC molecules perform a dual role of cell-surface marker and antigen-presenting structure (Figure 1). T cell accessory molecules recognize nonpolymorphic, class-specific MHC structures and are involved in determining the T cell effector response to antigens. Self-MHC restriction is primarily a consequence of the involvement of autologous MHC molecules in antigen presentation and their allele-specific contribution to interaction antigens. This model is based on a single-recognitive strategy for antigen/MHC recognition, but invokes additional receptors to facilitate this recognition and T cell activation.

Model Rationale

Structural Similarities Suggest that Important Features of Antigen Recognition Are Shared by B and T Cell Receptors

Several observations indicate that the secondary and tertiary structures of the V regions of α/β complexes and those of antibodies are fundamentally the same. First, they are similar in size and significantly homologous to one another. Each V region has two cysteines separated by 60 to 70 amino acids that define Ig homology units (Hood et al., 1985). Both V_α and V_β share essentially the same set of conserved residues involved in stabilizing the intra- and inter-chain structural features of Ig chains (for review, see Kronenberg et al., 1986). Second, when β-pleated sheet potential and the hydropathic profiles of T cell V regions are analyzed, the predicted structural features are remarkably similar to those of Ig V regions (Patten et al., 1984; Barth et al., 1985). These data suggest that T and B cell receptors fold into very similar tertiary structures based on an "antibody fold."

These structural similarities suggest that T and B cells use similar strategies for antigen recognition. Ig antigen-binding pockets are derived from three hypervariable sequences found on both L and H chains. Analyses of the sequence variability of V_α and V_β chains have been carried out by us and others (Patten et al., 1984; Arden et al., 1985; Barth et al., 1985; Becker et al., 1985; Behlke et al., 1985). We believe that the profiles of both are essentially the same as those of their Ig counterparts and reflect structurally similar binding pockets. In addition, antibodies can be selected that exhibit MHC-restriction comparable to that of T cell receptors (Wylie et al., 1982). This is consistent with observations indicating that B cell response to certain antigens can be MHC-restricted (Katz et al., 1978; Sprent and Bruce, 1979; Singer and Hodes, 1982). Thus, different requirements for activation may be more fundamental in distinguishing B and T cell responses to antigens than antigen recognition itself.

Expression of α and β Chain Gene Segments Correlates with Neither Antigen Nor MHC Specificity

Models of MHC-restricted T-cell antigen recognition differ by using either a single binding site for an antigen/MHC complex (single-recognitive), or separate, nonoverlapping sites specific for either antigen or MHC (dual-recognitive). A two-chain T cell receptor is consistent with a simple dual-recognitive model with V regions of α and β chains responsible for either MHC or antigen recognition exclusively. However, when pairs of T cells with different antigen and different MHC specificities are fused, they fail to generate the novel combinations of specificities expected if antigen and MHC are recognized independently (Kappler et al., 1981). In addition, there are several known examples of functional T cell clones that use the same V or J gene segments in either chain but have different antigen and MHC specificities (Goverman et al., 1985; Malissen et al., 1985; Rupp et al., 1985; Sim and Augustin, 1985). There are also examples of different T cell clones specific for the same antigen and MHC that use the same V_α or V_β gene segments but differ for the other chain. These results imply that MHC-restricted antigen specificity is a unique property of an α/β combination, and that the α and β chains together can contribute to both antigen and MHC recognition as L and H chains contribute to Ig antigen recognition.

MHC Polymorphism Plays an Important Role in Restricted Antigen Presentation and Ir Gene Effects

Two properties of MHC molecules impact on their role in antigen recognition: they are encoded in multiple, simultaneously expressed loci and they exhibit enormous polymorphism in populations. There are an estimated 50–100 expressed major alleles of class I and class II molecules in humans and mice (Klein, 1979). At least two class II loci in mice and three in humans are expressed (for review, see Mengle-Gaw and McDevitt, 1985). Allowing for combinatorial association, heterozygous individuals may express eight or more different class II molecules. Two to three class I loci are expressed in both mice and humans, with the potential expression of four to six antigen-presenting class I molecules (Bevan, 1981). A class I molecule is composed of one MHC-encoded heavy chain that associates with the non-MHC-encoded β_2-microglobulin. Class II molecules are composed of two chains, denoted α and β. Both class I and class II molecules appear to fold into similar, two-domain cell-surface structures (Kimball

and Coligan, 1983) (Figure 1). The membrane-proximal regions of both classes are C region-like homology units and structural analysis suggests that the distal regions may have a distant relationship to V homology units (Hood et al., 1983; Novotný and Auffray, 1984; Mengle-Gaw and McDevitt, 1985; Norcross and Kanehisa, 1985). The polymorphism of both class I and class II molecules is confined primarily to the V-like domains (for review, see Mengle-Gaw and McDevitt, 1985). Exon-shuffling experiments and analyses of mutant MHC polypeptides indicate that restricted antigen presentation and Ir effects segregate with these distal domains (Nairn et al., 1980; Evans et al., 1982; Sherman, 1982; Clark and Forman, 1983; Ozato et al., 1983; Reiss et al., 1983; Stroynowski et al., 1984).

Noncovalent associations have been observed between several combinations of class I molecules and particular viral proteins (for review, see Pääbo et al., 1985). Recently, Babbitt et al. (1985) demonstrated that an antigenic peptide of lysozyme binds to a class II molecule expressed by a responder, but not to one expressed by a nonresponder. Other work indicates that such interactions involve the polymorphic domain of class II molecules (for review, see Schwartz, 1985). These data suggest that antigen presentation may require direct binding of antigen by the MHC polymorphic domains. Consequently, we believe many Ir effects are due to the inability of an MHC allele to bind, and hence present, certain antigens. It is also possible that binding occurs but that MHC allele-specific contributions result in interaction complexes for which no responding T cells can be generated. Such "holes" in the repertoire could result from genetic deficiencies of certain gene segments or from specific tolerance requirements. Our model is consistent with both explanations for Ir effects, as both have a basis in direct antigen-MHC interaction. The difficulty in demonstrating such interaction in vitro may arise from the fact that particular antigen-MHC affinity must be low in order for a limited number of MHC molecules to present a broad range of antigens. Interaction with the α/β receptor may further stabilize some complexes (Schwartz, 1985).

Another consequence of direct antigen-MHC interaction postulated by interaction-antigen models is that one antigen/self-MHC pair can generate the same local combinatorial determinants as a different antigen/allo-MHC pair (i.e., self-MHC plus antigen X can equal allo-MHC plus antigen Y) and, hence, can appear to be the same antigenic structure to a particular T cell. Hünig and Bevan (1982) described a T cell clone specific for two separate minor histocompatibility antigens, each restricted by different, nonallelic class I molecules. Reciprocal combinations of antigen and MHC were not recognized by the T cell clone, indicating that each specificity was for a complex rather than for the individual components. Similarly, several "degenerate" T cell clones have been isolated which recognize the same antigens, or only slightly altered antigens, when presented by allelic MHC products (for review, see Schwartz, 1985). The same MHC molecules can present other antigens such that their recognition is restricted exclusively of the other MHC alleles. This demonstrates that this degeneracy is also a consequence of the antigen-MHC combinations and not simple cross-reactivity.

Alloreactivity May Be the Same Phenomenon as Degenerate Antigen Recognition

Antigen-specific T cell clones have a 2%–5% chance of recognizing any single allo-MHC tested, indicating that alloreactive T cells are not a separate subset of the normal T cell population (Schwartz and Sredni, 1982). Both molecular and cellular data seem to rule out the idea that these cells have separate receptors specific for only allo-MHC (von Boehmer et al., 1979; Sredni and Schwartz, 1981; Malissen et al., 1985; Yagüe et al., 1985). If T cell receptors do not distinguish the source of interaction-antigen determinants, one receptor could recognize both a specific antigen/MHC complex and an allo-MHC molecule alone, provided that its structure mimics the novel determinants of the complex. In support of this, recent work indicates that the antigen specificity of a T cell clone biases its specificity for particular allo-MHC molecules (Ashwell et al., 1986). Our model predicts a bias in the peripheral population for such cross-reactivity because the polymorphic domains of different MHC alleles already share many determinants involved in interaction-antigens. Therefore, we believe that many allogeneic responses represent cross-reactivity for interaction-antigen complexes. Alloreactivity may be amplified because the concentration of MHC molecules on any single cell will generally be much greater than that of presented antigen. Alloreactivity may also be due to allo-presentation of either non-MHC allo-antigens or self-antigens to which the T cell population is tolerant in the context of self-MHC molecules.

MHC Polymorphism Functionally Expands the T Cell Receptor Repertoire

In dual-recognitive models, T cells require separate pools of diversity to accommodate both MHC molecules and antigen (Pernis and Axel, 1985). Thus, minimun T cell diversity must equal the number of potential antigens multiplied by the number of MHC alleles. Interaction-antigen models are not constrained by this combinatorial because the determinants of an interaction complex cannot be distinguished from those found in the general pool of potential antigens. We believe that MHC polymorphism in fact reduces the number of T cell receptors necessary to recognize a given pool of antigens. Expression of multiple MHC molecules provides the potential for presenting an antigen in different contexts, thereby increasing the likelihood that the T cell repertoire has a complementary receptor structure. Extensive polymorphism insures this hetergeneity in individuals.

Our view predicts that the degree of MHC polymorphism expressed in individuals and populations should reflect the potential antigenic exposure. The Syrian hamster is essentially nonpolymorphic for class I molecules, which usually present viral antigens, but is polymorphic for class II molecules (Streilein and Duncan, 1983). It was suggested that the Syrian hamster is seldom exposed to

viruses due to its extremely solitary lifestyle and thus may have a reduced need for diverse structures to present viral antigens. In a complementary manner, sudden loss of MHC polymorphism within a population exposed to a normal antigenic environment should place that population at increased risk. Cheetahs seem particularly susceptible to viral epizootics relative to other populations of big cats. Unlike other big cats, cheetahs lack polymorphism for class I, presumably because of a recent founder event resulting from overhunting (O'Brien et al., 1985).

Recognition of MHC Class Correlates with T Cell Effector Function and Involves Sequences Unrelated to Those Involved in Antigen Recognition

Different Ig effector functions are mediated by different C_H regions or isotypes. The expression of the two C_β gene segments does not correlate with differences in T cell function and there is also no indication that gene segments of either chain divide into class-specific subsets (Becker et al., 1985). If there is no restriction on combinatorial association between V_α and V_β genes, it seems unlikely that particular combinations would generate class-specific receptor sites. Therefore, the α/β receptor is insufficient to define T cell functional response.

Because the class of MHC molecule presenting antigen correlates well with the effector function of responding T cells, recognition of class-specific determinants may help define functional response. It has been suggested that MHC class is recognized independently of allele-specific determinants by molecules other than the antigen-specific receptor (Biddison et al., 1982; Swain et al., 1983). Experiments in which T_C function was inhibited by anti-Lyt2 antibodies led to the proposal that Lyt2 directly recognizes class I MHC products (Wettstein et al., 1978; Okada and Henney, 1980; Swain, 1981) (Figure 1). Similarly, T4 and L3T4 may recognize class II molecules (Biddison et al., 1982; Marrack et al., 1983). The presence of Lyt2 and L3T4 is more strictly correlated with class I and class II antigen presentation, respectively, than with particular T cell functions. The V-like structure of these molecules suggests that they function as receptors for MHC class-specific determinants. However, following in vivo immunization it is possible to isolate class I-restricted T cells that respond to antigen in vitro even when Lyt2 is blocked by antibody (MacDonald et al., 1981). It was proposed that these cells were selected in vivo for high antigen avidity and that Lyt2 and L3T4 operate to enhance the overall avidity between T cells and target cells.

There are various mechanisms by which Lyt2 and L3T4 could modulate cell–cell avidity. By binding to MHC molecules, they could bring T cell and target cell surfaces together transiently and increase the likelihood that an α/β complex will encounter only presented antigen. In addition to shifting the kinetics of cellular interactions, accessory molecule–MHC binding could increase the avidity of target and effector cells and compensate for lower avidity α/β-antigen/MHC interactions or for low concentration of presented antigen. This process could occur between any target cell and T cell with the appropriate combination of MHC and accessory molecule, regardless of whether antigen was present. In this scheme, α/β receptors could bind antigen presented by any MHC molecule—not just those MHC molecules directly bound by the accessory molecules. This implies that, because APCs expressing class II molecules also express class I products, both Lyt2 and L3T4 would facilitate interaction of an α/β complex with a class II MHC/antigen complex (Figure 2). Therefore, the correlation between Lyt2 expression and recognition of class I-presented antigen should be less exclusive than the correlation between L3T4 expression and recognition of class II-presented antigen. Because breakdown of the specificities of both these accessory molecules is rarely observed, it may be more likely that Lyt2 and L3T4 specifically direct α/β complexes toward the same MHC molecules bound by the accessory molecules. We have described a six residue (6-mer) sequence conserved between the polymorphic regions of class I molecules and those of the class II β-chains, and suggested that this sequence might be part of a general target for T cell recognition of MHC molecules (Malissen et al., 1983). A related but distinct peptide is also conserved in class II α-chains, but not in the structurally equivalent region of class I molecules. This α-chain peptide, in conjunction with class-specific variation in the sequences neighboring the 6-mer, may help functionally define MHC class. T cell response to xenogeneic class I molecules and recent results of experiments with class II/class I chimeric molecules suggest that both Lyt2 and L3T4 interact with the polymorphic domain of MHC molecules (Swain et al., 1983; Golding et al., 1985a). We propose that Lyt2 and L3T4 are specific for 6-mer-related sequences and bias T cell surveillance toward the antigen-presenting domain of the same target cell MHC molecules. Some Lyt2 and L3T4 blocking experiments indicate that these molecules are not always required for T cell antigen recognition. However, the high concentration of antigen used in in vitro experiments relative to that encountered in vivo may decrease the need for

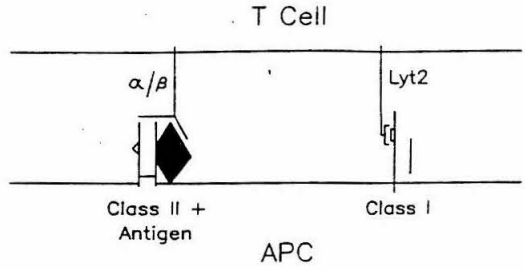


Figure 2. A Possible Mode of T Cell/Target Cell Interaction Mediated by Lyt2

MHC class II-expressing APCs also express class I molecules. This figure illustrates the interaction of an Lyt2-expressing T cell with a target cell that is presenting antigen (black diamond) with a class II molecule. If the recognition functions of the α/β receptor and the accessory molecules (Lyt2 or L3T4) are not linked, Lyt2 could bind to an MHC molecule not involved in antigen presentation. Therefore, Lyt2 should be able to mediate T cell recognition of an MHC class II-presented antigen. This scenario predicts a less exclusive correlation between L3T4 expression by T cells and recognition of class II-presented antigens.

MHC targeting by increasing the number of interaction complexes on the cell surface. As a refinement of this directed scheme, accessory molecules might preferentially direct α/β complexes toward MHC molecules that are conformationally altered by antigen presentation. Pääbo et al. (1985) suggest that the 6-mer may be more exposed when the MHC molecule is sterically modified, perhaps by antigen interaction. Whether or not the process is antigen independent, direction of the α/β receptor by Lyt2 and L3T4 requires a testable interaction between these molecules and the receptor.

T Cell Activation

The α/β complex in our model does not distinguish MHC from antigen determinants and consequently could bind free antigen without other constraints. We believe accessory molecules direct T cell receptors to MHC molecules and, hence, could convey at least some of these constraints. Since some T cells can be stimulated in the absence of accessory molecule interaction, the role of these molecules in T cell activation cannot be essential in all situations (Marrack et al., 1983). T cells can also be activated in vitro independently of antigen recognition by various lectins, lymphokines, and serological reagents directed toward different T cell surface structures. Therefore, T cell activation in vivo probably involves several different components (Greenstein et al., 1984). In our model, antigen recognition is required but not sufficient for T cell activation. Additional requirements may be either transmission of other specific signals or formation of an activation complex in which multiple molecules coordinately transmit an antigen-recognition signal. In addition to Lyt2 and L3T4, there are numerous other cell surface molecules that appear to be important in T cell activation (for review, see Townsend, 1985). Different subsets of these molecules may participate in activation depending on the conditions of antigen presentation. For example, different types of APCs may interact with particular subsets of these molecules.

T Cell Ontogeny Results in But Does Not Select for a Population Biased for Recognition of Antigen Presented by Self-MHC

Although our model is primarily functional, it must be viewed in a broader developmental context. In particular, our model has implications for the process of thymic education. Most dual-recognitive models have been proposed specifically to explain positive thymic selection for self-MHC-restricted T cells. However, while some bone marrow chimeras appear to demonstrate nearly complete restriction of the donor T cells to host MHC molecules, many similar experiments demonstrate only a bias (Blanden and Andrew, 1979; for review, see Klein and Nagy, 1982). Most studies of MHC-restriction in normal animals also do not show absolute restriction of T cells to thymic MHC molecules. If a normal T cell population is first depleted of alloreactive clones by negative-selection techniques or by rendering mice neonatally tolerant to an allo-MHC molecule, a portion of the remaining T cells can respond to an antigen presented by that allo-MHC molecule (for review, see Schwartz, 1984). In several experiments, these allo-restricted cells were shown to be a separate population from those cells that were self-restricted for the same antigen. Although some negative-selection experiments did not indicate allo-MHC antigen presentation, most experiments compared self-restricted response to the response for only one or two other allo-MHC alleles. The results, therefore, seem to depend on the particular combinations of antigen and MHC alleles tested (Doherty and Bennink, 1979). The actual contribution of the thymus to MHC-restriction has been investigated by determining the ratio of self- to allo-restricted T cells responding to a particular antigen in T cell populations that had not been primed previously with that antigen (Wagner et al., 1981; Stockinger et al., 1981). In general, the ratio was found to be only about 2–6 to 1 for any one allo-MHC tested in both allogeneic chimeras and normal mice, with no differences between thymic and splenic T cell populations. When all these experiments are considered, a clear pattern emerges: restriction to thymic MHC type is not functionally absolute, but is only a bias. This implies that models of T cell antigen recognition are not constrained to accommodate strict selection for thymic MHC type and that T cells exiting the thymus express a receptor repertoire essentially unlimited in its ability to recognize antigen with allo-MHC molecules.

In Figure 3 we illustrate a program of T cell maturation consistent with our view of T cell antigen recognition. Interaction of the α/β receptor with some ligand might be important at one or more stages of development to facilitate or redirect this program. These stages may be defined by susceptibilities to various lymphokines, perhaps present in different developmental compartments of the thymus (for review, see Reinherz, 1985). If a T cell were to be triggered at an early stage by binding a self-MHC molecule or a self-MHC/antigen interaction complex, it would be eliminated or made tolerant. A similar stimulus at a later stage, on the other hand, may trigger functional maturation. It is possible that T cells must be tested for the presence of a functional receptor during this later stage before exit to the periphery. This process would favor those T cells that interact with ligand, but others may passively "escape" from the thymus by some stochastic process. Two-thirds or more of developing T cells may not produce a full complement of functional receptor genes because of aberrant rearrangements (Kronenberg et al., 1985), and therefore would fail a test for a functional receptor. This attrition, and that caused by the elimination of cells for self-reactivity, could account for a great deal of the observed cell turnover in the thymus.

We believe that antigen priming in the context of self-MHC alleles is primarily responsible for the bias toward self-MHC-restricted T cells in the periphery of normal animals. However, the thymus may be the first place maturing T cells have an opportunity to encounter antigen. Longo and Schwartz (1980) have shown that T cells recognizing class II-presented antigen are restricted primarily to the MHC molecules of thymic APCs rather than to those of thymic epithelial cells. The involvement of APCs in this restriction suggests to us a role for self-MHC

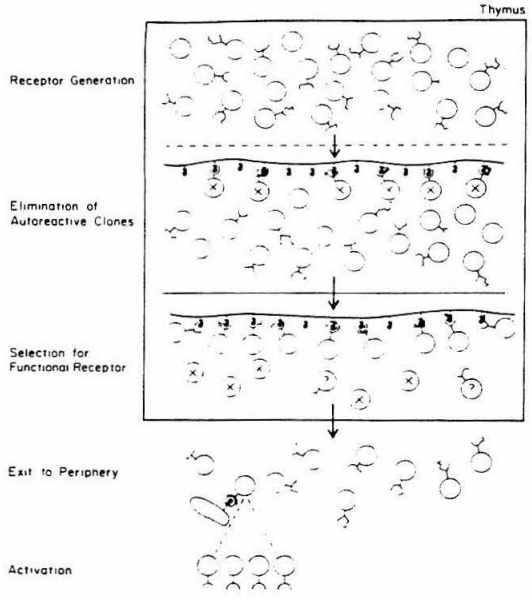


Figure 3. Simple Model of T Cell Repertoire Development in the Thymus

After T cells with unique receptors are generated, there is a negative selection against those that are complementary to self-MHC molecules alone (small filled forms attached to a wavy cell membrane), or to self-MHC molecules presenting antigen (open forms attached to MHC molecules). An X indicates that the responding T cell will be eliminated. A dashed line between the first two phases indicates that there may be a compartmental or temporal separation between these two events, but that such a distinction is not required. Once autoreactive clones have been eliminated or made tolerant, remaining clones are tested for the presence of a functional receptor by interaction with MHC-presented antigen or by low-affinity, nonspecific interactions with self-MHC molecules. Presumably, those T cells without a receptor complex are not triggered to leave the thymus. Those T cells with otherwise functional receptors that are not triggered sufficiently by the expressed MHC molecules may be eliminated; alternatively, they may escape through a more stochastic process, albeit at a lower frequency than stimulated T cells. A solid line indicates that this later process must be separated from the negative selection process. After exit from the thymus, the surviving T cells are available for activation and proliferation.

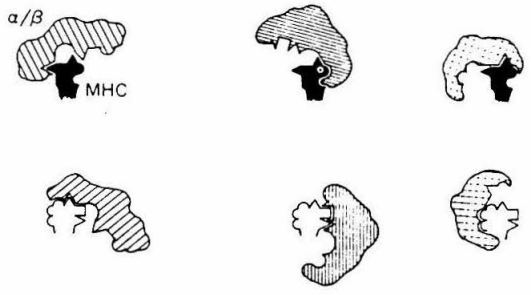


Figure 4. Low-affinity, Nonspecific Interaction between T Cell Receptors and Self-MHC Molecules in the Thymus

Identical α/β receptors have the same hatching pattern. Identical MHC molecules are filled or not filled similarly. The same T cell receptor is shown interacting with nonhomologous determinants on different MHC molecules in the thymus. Also included are different T cell receptors interacting with the same MHC molecule. This illustration is designed to indicate that there is no requirement for an MHC-specific recognition region on a T cell receptor or a specifically recognized structure on the MHC molecule to impart a bias for self-MHC recognition in a T cell repertoire.

plus antigen in biasing the T cell repertoire. Although proliferative response of thymocytes to antigen is difficult to detect (for review, see Rothenberg and Lugo, 1985), antigen priming in the thymus can be demonstrated (Fink et al., 1984; Kyewski et al., 1984). Other experiments fail to show any involvement of APCs with class I-restricted T cells (Zinkernagel, 1982). However, recognition of class I-presented antigen is generally less restricted to thymic MHC alleles than is recognition of class II-presented antigen (Bradley et al., 1982). Recent experiments suggest that this distinction may reflect differences in developmental influences on T_H and T_C cell subsets (Kruisbeek et al., 1983; Golding et al., 1985b). For example, the stronger influence of the thymus on class II-restricted T cells may be due in part to a greater likelihood of class II-presented antigens being found in the thymus.

Our view of the T cell receptor suggests that receptor interactions in the thymus would frequently involve antigen/self-MHC complexes. However, low-affinity binding of some thymocytes to self-MHC molecules alone may also be sufficient for stimulating maturation (Reinherz, 1985). In our model, this interaction would not be mediated by a specific binding site for MHC and would not necessarily involve allele-specific determinants. Therefore, it does not reflect a required selection for recognition of self-MHC. The receptors must merely accommodate the MHC molecules present, not be specific for them (Figure 4). The MHC determinants that would participate in these low-affinity interactions are presumably the same determinants involved in antigen interactions and would often be altered or lost in particular interaction complexes. However, complexes would form in which some of the original MHC determinants remained. A T cell population expanded through this process should express a bias for recognition of interaction antigens that retain original MHC determinants. This process may explain why T cell recognition of smaller antigens is generally more restricted than recognition of larger ones; that is, smaller antigens would disrupt or mask the MHC determinants to a lesser degree than larger antigens (Schwartz, 1984).

Members of the Ig Superfamily Use Similar Recognitive Strategies

T cell receptors, class I and class II MHC molecules, and accessory T cell molecules such as Lyt2 and L3T4 all belong to the Ig superfamily. We suggest that these molecules use specific recognition sites structurally homologous to the Ig antigen binding pocket. The structural and functional potentials of these V-like regions are the consequence of two fundamental properties of homology

units. First, they tend to form stable, pairwise associations (domains) with other similar units. The binding pocket is essentially the cleft formed between the two interacting V-like regions; hence, both chains are able to contribute to specificity. Second, with conservation of only about fifteen specific amino acids, most of the V-like sequence is relatively free of evolutionary constraints. The resulting potential for diversity accounts for the exquisite specificity and range of recognition structures available to the vertebrate immune system. Given these two principles, the simplest scenarios for the evolution of the Ig superfamily do not require radical variation on the basic receptor structure, but rather require use of the same general recognitive strategies in novel contexts.

Conclusion

We have presented here what we believe to be a simple model of T cell response to antigen and have attempted to integrate the growing body of molecular data with the rich body of information derived from cellular research. Other models have recently been proposed that also take into account aspects of the molecular data, but, unlike ours, accommodate positive selection for self-MHC (Schwartz, 1985; Norcross and Kanehisa, 1985; Langman and Cohn, 1985; Pernis and Axel, 1985; Raulet et al., 1985). We have presented our model in the hope that its testing will contribute to a more complete understanding of the immune response.

Acknowledgments

We thank Drs. U. Landegren, J. Lugo, E. Rothenberg, and I. Weissman for helpful and stimulating discussions; Drs. S. Hunt, M. Kronenberg, and N. Shastri for comments on the manuscript; and C. Elkins and C. Katz for tireless efforts in preparing the manuscript. The estimate of the number of V_H gene segments in Table 1 is based on information kindly supplied by Dr. D. Livant. J. G. is a Special Fellow of the Leukemia Society of America.

References

Acuto, O., Meuer, S., Hodgdon, J., Schlossman, S., and Reinherz, E. (1983). Peptide variability exists within α and β subunits of the T-cell receptor for antigen. J. Exp. Med. *158*, 1368–1373.

Allison, J., McIntyre, B., and Block, D. (1982). Tumor-specific antigen of murine T lymphoma defined with monoclonal antibody. J. Immunol. *129*, 2293–2300.

Arden, B., Klotz, J., Siu, G., and Hood, L. E. (1985). Diversity and structure of genes of the α family of mouse T-cell antigen receptor. Nature *316*, 783–787.

Ashwell, J. D., Chen, C., and Schwartz, R. H. (1986). High frequency and nonrandom distribution of alloreactivity in T-cell clones selected for recognition of foreign antigen in association with self-class II molecules. J. Immun. *136*, 389–395.

Babbitt, B., Allen, P. M., Matsudeda, G., Haber, E., and Unanue, E. R. (1985). Binding of immunogenic peptides to Ia histocompatibility molecules. Nature *317*, 359–361.

Barth, R., Kim, B., Lan, N., Hunkapiller, T., Sobieck, N., Winoto, A., Gershenfeld, H., Okada, C., Hansburg, D., Weissman, I., and Hood, L. (1985). The murine T-cell receptor uses a limited repertoire of expressed V_β gene segments. Nature *316*, 517–523.

Becker, D., Patten, P., Chien, Y.-H., Yokota, T., Eshhar, Z., Giedlin, M., Gascoigne, N. R. J., Goodenow, C., Wolf, R., Arai, K.-I., and Davis, M. M. (1985). Variability and repertoire size of T-cell receptor V_α gene segments. Nature *317*, 430–434.

Behlke, M. A., Spinella, D. G., Chou, H., Sha, W., Hartl, D. L., and Loh, D. Y. (1985). T-cell receptor β chain expression: dependence on relatively few variable region genes. Science *229*, 566–570.

Bevan, M. J. (1981). Thymic education. Immunol. Today *2*, 216–219.

Biddison, W. E., Rao, P. E., Talle, M. A., Goldstein, G., and Shaw, S. (1982). Possible involvement of the OKT4 molecule in T-cell recognition of class II HLA antigens. J. Exp. Med. *156*, 1065–1076.

Blanden, R. V., and Andrew, M. E. (1979). Primary anti-viral cytotoxic T-cell responses in semiallogeneic chimeras are not absolutely restricted to host H-2 type. J. Exp. Med. *149*, 535–538.

Bradley, S. M., Kruisbeek, A. M., and Singer, A. (1982). Cytotoxic T lymphocyte response in allogeneic radiation bone marrow chimeras. J. Exp. Med. *156*, 1650–1664.

Chien, Y., Becker, D. M., Lindsten, T., Okamura, M., Cohen, D. I., and Davis, M. M. (1984). A third type of murine T-cell receptor gene. Nature *312*, 31–35.

Clark, S. S., and Forman, J. (1983). Allogeneic and associative recognition determinants of H-2 molecules. Transplant. Proc. *15*, 2090–2092.

Davis, M. M. (1985). Molecular genetics of the T cell-receptor beta chain. Ann. Rev. Immunol. *3*, 537–560.

Doherty, P. C., and Bennink, J. R. (1979). Patterns of virus-immune T-cell responsiveness. J. Exp. Med. *150*, 1187–1194.

Doherty, P. C., and Zinkernagel, R. M. (1975). H-2 compatibility is required for T-cell mediated types of target cells infected with lymphocytic choriomeningitis virus. J. Exp. Med. *141*, 502–507.

Evans, G. A., Margulies, D. H., Ozato, K., Camerini-Otero, R. D., Ozato, K., and Seidman, J. G. (1982). Structure and expression of a mouse major histocompatibility antigen gene, H-2L^d. Proc. Natl. Acad. Sci. USA *79*, 1994–1998.

Fink, P. J., and Bevan, M. J. (1978). H-2 antigens of the thymus determine lymphocyte specificity. J. Exp. Med. *148*, 766.

Fink, P. J., Bevan, M. J., and Weissman, I. J. (1984). Thymic cytotoxic T lymphocytes are primed in vivo to mirror histocompatibility antigen. J. Exp. Med. *159*, 436–451.

Fischer-Lindahl, K., and Wilson, D. B. (1977). Histocompatibility antigen-activated cytotoxic T lymphocytes. II. Estimates of the frequency and specificity of precursors. J. Exp. Med. *145*, 508–522.

Ford, W. L., Simmonds, S. J., and Atkins, R. C. (1975). Early allelic events in a systemic graft-vs-host reaction II. Autoradiographic estimates of the frequency of donor lymphocyte which respond to each Ag-B-determined antigenic complex. J. Exp. Med. *141*, 681–696.

Gascoigne, N., Chien, Y., Becker, D., Kavaler, J., and Davis, M. (1984). Genomic organization and sequence of T-cell receptor β-chain constant- and joining-region genes. Nature *310*, 387–391.

Gearhart, P. J., Johnson, N. D., Douglas, R., and Hood, L. (1981). IgG antibodies to phosphorylcholine exhibit more diversity than their IgM counterparts. Nature *291*, 29–34.

Golding, H., McCluskey, J., Munitz, T. I., Germain, R. N., Margulies, D. H., and Singer, A. (1985a). T-cell recognition of a chimaeric class II/class I MHC molecule and the role of L3T4. Nature *317*, 425–427.

Golding, H., Munitz, T. I., and Singer, A. (1985b). Characterization of antigen-specific, Ia-restricted, L3T4 cytolytic T lymphocytes and assessment of thymic influence on their self-specificity. J. Exp. Med. *162*, 943–961.

Goverman, J., Minard, K., Shastri, N., Hunkapiller, T., Hansburg, D., Sercarz, E., and Hood, L. (1985). Rearranged β T cell receptor genes in a helper T cell clone specific for lysozyme: no correlation between V_β and MHC restriction. Cell *40*, 859–867.

Greenstein, J., Kappler, J., Marrack, P., and Burakoff, S. (1984). The role of L3T4 in recognition of Ia by a cytotoxic H-2D^d-specific T-cell hybridoma. J. Exp. Med. *159*, 1213–1224.

Haskins, K., Kubo, R., White, J., Pigeon, M., Kappler, J., and Marrack, P. (1983). The major histocompatibility complex-restricted antigen receptor in T cells. I. Isolation with a monoclonal antibody. J. Exp. Med. *157*, 1149–1169.

Hayday, A. C., Diamond, D. J., Tanigawa, G., Heilig, J. S., Folsom, V., Saito, H., and Tonegawa, S. (1985a). Unusual organization and diversity of T-cell receptor α-chain genes. Nature *316*, 828–832.

Hayday, A. C., Saito, H., Gillies, S. D., Kranz, D. M., Tanigawa, G., Eisen, H. N., and Tonegawa, S. (1985b). Structure, organization, and somatic rearrangement of T cell gamma genes. Cell *40*, 259–269.

Herron, J. N., and Voss, E. W., Jr. (1983). Analysis of heterogeneous dissociation kinetics in polyclonal populations of rabbit anti-fluorescyl-IgG antibodies. Molecular Immunology *20*, 1323–1332.

Hood, L., Kronenberg, M., and Hunkapiller, T. (1985). T cell antigen receptors and the immunoglobulin supergene family. Cell *40*, 225–229.

Hood, L., Steinmetz, M., and Malissen, B. (1983). Genes of the major histocompatibility complex of the mouse. Ann. Rev. Immunol. *1*, 529–568.

Hünig, T. R., and Bevan, M. J. (1982). Antigen recognition by cloned cytotoxic T lymphocytes follows rules predicted by the altered-self hypothesis. J. Exp. Med. *155*, 111–125.

Kappler, J. W., and Marrack, P. (1976). Helper T cells recognize antigen and macrophage surface components simultaneously. Nature *262*, 797–799.

Kappler, J., Kubo, R., Haskins, K., Hannum, C., Marrack, P., Pigeon, M., McIntyre, B., Allison, J., and Trowbridge, I. (1983). The major histocompatibility complex-restricted antigen receptor on T cells in mouse and man: identification of constant and variable peptides. Cell *35*, 295–302.

Kappler, J. W., Skidmore, B., White, J., and Marrack, P. (1981). Antigen-inducible, H-2 restricted interleukin-2-producing T-cell hybridomas: lack of independent antigen and H-2 recognition. J. Exp. Med. *153*, 1198–1214.

Katz, D. H., Hamaoka, T., Dorf, M. E., and Benacerraf, B. (1973a). Cell interactions between histoincompatible T and B lymphocytes: the H-2 gene complex determines successful physiologic lymphocyte interactions. Proc. Natl. Acad. Sci. USA *70*, 2624–2628.

Katz, D. H., Hamaoka, T., Dorf, M. E., Maurer, P. H., and Benacerraf, B. (1973b). Cell interactions between histoincompatible T and B lymphocytes. IV. Involvement of the immune response (Ir) gene in the control of lymphocyte interactions in responses controlled by the gene. J. Exp. Med. *138*, 734–739.

Katz, D. H., Skidmore, B. J., Katz, L. R., and Bigowitz, C. A. (1978). Adaptive differentiation of murine lymphocytes. I. Both T and B lymphocytes differentiating in $F_1 \rightarrow$ parental chimeras manifest preferential cooperative activity for partner lymphocytes derived from the same parental type corresponding to the chimeric host. J. Exp. Med. *148*, 727–745.

Kimball, E. S., and Coligan, J. E. (1983). Structure of class I major histocompatibility antigens. Contemp. Topics Molec. Immunol. *9*, 1–63.

Kindred, B., and Shreffler, D. C. (1972). H-2 dependence of cooperation between T and B cells *in vivo*. J. Immunol. *109*, 940–943.

Klein, J. (1979). The major histocompatibility complex of the mouse. Science *203*, 516–521.

Klein, J., and Nagy, Z. A. (1982). MHC restriction and Ir genes. Adv. Cancer Res. *37*, 234–317.

Kranz, D. M., Saito, H., Heller, M., Takagaki, Y., Haas, W., Eisen, H. N., and Tonegawa, S. (1985). Limited diversity of the rearranged T-cell γ gene. Nature *313*, 752–955.

Kronenberg, M., Goverman, J., Haars, R., Malissen, M., Kraig, E., Phillips, L., Delovitch, T., Suciu-Foca, N., and Hood, L. (1985). Rearrangement and transcription of the β-chain genes of the T-cell antigen receptor in different types of murine lymphocytes. Nature *313*, 647–653.

Kronenberg, M., Siu, G., Hood, L., and Shastri, N. (1986). The molecular genetics of the T-cell antigen receptor and T-cell antigen recognition. Ann. Rev. Immunol., in press.

Kruisbeek, A. M., Sharrow, S. O., and Sanger, A. (1983). Differences in the MHC-restricted self-recognition repertoire of intra-thymic and extra-thymic cytotoxic T lymphocyte precursors. J. Immunol. *130*, 1027–1032.

Kyewski, B. A., Fathman, C. G., and Kaplan, H. S. (1984). Intrathymic presentation of circulating non-major histocompatibility complex antigens. Nature *308*, 196–199.

Langman, R. E., and Cohn, M. (1985). T cells function via restricted recognition of antigen, not antigen-restricted recognition. Cell. Immunol. *94*, 598–608.

Littman, D., Thomas, Y., Maddon, P., Chess, L., and Axel, R. (1985). The isolation and sequence of the gene encoding T8: a molecule defining functional classes of T lymphocytes. Cell *40*, 237–246.

Longo, D. L., Matis, L. A., and Schwartz, R. H. (1981). Insights into immune response gene function from experiments with chimeric animals. Crit. Rev. Immunol. *2*, 83–132.

Longo, D. L., and Schwartz, R. H. (1980). T-cell specificity for H-2 and Ir gene phenotype correlates with the phenotype of thymic antigen-presenting cells. Nature *287*, 44–46.

Luzzati, A. L., Lefkovits, I., and Pernis, B. (1973). Homogeneity of antibodies produced by clones in vitro. Eur. J. Immunol. *3*, 636–639.

MacDonald, H. R., Thiernesse, N., and Cerottini, J. (1981). Inhibition of T-cell mediated cytolysis by monoclonal antibodies directed against Lyt-2: heterogeneity of inhibition at the clonal level. J. Immunol. *126*, 1671–1675.

Maddon, P., Littman, P., Godfrey, M., Maddon, D., Chess, L., and Axel, R. (1985). The isolation and nucleotide sequence of a cDNA encoding the T cell surface protein T4: a new member of the immunoglobulin gene family. Cell *42*, 93–104.

Malissen, M., Hunkapiller, T., and Hood, L. (1983). Nucleotide sequence of a light chain gene of the mouse I-A subregion: A^q_β. Science *221*, 750–754.

Malissen, M., Minard, K., Mjolsness, S., Kronenberg, M., Goverman, J., Hunkapiller, T., Prystowsky, M., Yoshikai, Y., Fitch, F., Mak, T., and Hood, L. (1984). Mouse T cell antigen receptor: structure and organization of constant and joining gene segments encoding the β polypeptide. Cell *37*, 1101–1110.

Malissen, M., McCoy, C., Blanc, P., Trucy, J., Devaux, C., Schmitt-Verhulst, A.-M., Fitch, F., Hood, L., and Malissen, B. (1986). Direct evidence for chromosomal inversion during T-cell receptor β gene rearrangements. Nature *319*, 28–33.

Marrack, P., Endres, R., Shimonkevitz, R., Zlotnick, A., Dialynas, D., Fitch, F., and Kappler, J. (1983). The major histocompatibility complex-restricted antigen receptor on T cells. II. Role of the L3T4 product. J. Exp. Med. *158*, 1077–1091.

McIntyre, B., and Allison, J. (1983). The mouse T cell receptor: structural heterogeneity of molecules of normal T cells defined by xenoantiserum. Cell *34*, 739–746.

Mengle-Gaw, L., and McDevitt, H. O. (1985). Genetics and expression of mouse Ia antigens. Ann. Rev. Immunol. *3*, 367–396.

Meuer, S. C., Acuto, O., Hussey, R. E., Hodgdon, J. C., Fitzgerald, K. A., Schlossman, S. F., and Reinherz, E. L. (1983). Evidence for the T3-associated 90 K heterodimer as the T-cell antigen receptor. Nature *303*, 808–810.

Meuer, S. C., Schlossman, S. F., and Reinherz, E. L. (1982). Clonal analysis of human cytotoxic T lymphocytes: $T4^+$ and $T8^+$ effector T cells recognize products of different major histocompatibility complex regions. Proc. Natl. Acad. Sci. USA *79*, 4395–4399.

Nairn, R., Yanoga, K., and Nathenson, S. G. (1980). Biochemistry of the gene products from murine MHC mutants. Ann. Rev. Genet. *14*, 241–277.

Norcross, M. A., and Kanehisa, M. (1985). The predicted structure of the Ia β_1 domain: a hypothesis for the structural basis of major histocompatibility complex-restricted T-cell recognition of antigens. Scand. J. Immunol. *21*, 511–524.

Novotný, J., and Auffray, C. (1984). A program for prediction of protein secondary structure from nucleotide sequence data: application to histocompatibility antigens. Nucl. Acids Res. *12*, 243–255.

O'Brien, S. J., Roelke, M. E., Marker, L., Newman, A., Winkler, C. A., Meltzer, D., Colly, L., Evermann, J. F., Bush, M., and Wildt, D. E. (1985). Genetic basis for species vulnerability in the cheetah. Science *227*, 1428–1434.

Okada, M., and Henney, C. S. (1980). The differentiation of cytotoxic T cells *in vitro*. II. Amplifying factor(s) produced in primary mixed lymphocyte cultures against K/D stimuli require the presence of Lyt2$^+$ cells but not Lyt1$^+$ cells. J. Immunol. *125*, 300–307.

Ozato, K., Evans, G. A., Shykind, B., Margulies, D., and Seidman, J. G. (1983). Hybrid H-2 histocompatibility gene products assign domains recognized by alloreactive T cells. Proc. Natl. Acad. Sci. USA *80*, 2040–2043.

Pääbo, S., Kämpe, O., Severinsson, L., Andersson, M., Fernandez, C., and Peterson, P. A. (1985). The association between class I transplantation antigens and an adenovirus membrane protein. Prog. Allergy *36*, 114–134.

Patten, P., Yokota, T., Rothbard, J., Chien, Y., Arai, K., and Davis, M. (1984). Structure, expression and divergence of T-cell receptor β-chain variable regions. Nature *312*, 40–46.

Pernis, B., and Axel, R. (1985). A one and a half receptor model for MHC-restricted antigen recognition by T lymphocytes. Cell *41*, 13–16.

Raulet, D. H., Gauman, R. D., Saito, H., and Tonegawa, S. (1985). Developmental regulation of T-cell receptor gene expression. Nature *314*, 103–107.

Reiss, C., Evans, G. A., Margulies, D. H., Seidman, J. G., and Burakoff, S. J. (1983). Allospecific and virus-specific cytolytic T lymphocytes are restricted to the N or C1 domain of H-2 antigens expressed on L cells after DNA-mediated gene transfer. Proc. Natl. Acad. Sci. USA *80*, 2709–2712.

Reinherz, E. L., (1985). A molecular basis for thymic selection: regulation of T11 induced thymocyte expansion by the T3-Ti antigen/MHC receptor pathway. Immunol. Today *6*, 75–79.

Rothenberg, E., and Lugo, J. P. (1985). Differentiation and cell division in the mammalian thymus. Dev. Biol. *112*, 1–17.

Rupp, F., Acha-Orbea, H., Hemgartner, H., Zinkernagel, R., and Joho, R. (1985). Identical V_β T-cell receptor genes used in alloreactive cytotoxic and antigen plus I-A specific helper T cells. Nature *315*, 425–427.

Saito, H., Kranz, D. M., Takagaki, Y., Hayday, A. C., Eisen, H. N., and Tonegawa, S. (1984a). A third rearranged and expressed gene in a clone of cytotoxic T lymphocytes. Nature *312*, 36–40.

Saito, H., Kranz, D., Takagaki, Y., Hayday, A., Eisen, H., and Tonegawa, S. (1984b). Complete primary structure of a heterodimeric T-cell receptor deduced from cDNA sequences. Nature *309*, 757–762.

Schwartz, R. H. (1984). The role of gene products of the major histocompatibility complex in T-cell activation. In Fundamental Immunology, W. E. Paul, ed. (New York: Raven Press), pp. 379–438.

Schwartz, R. H. (1985). T lymphocyte recognition of antigen in association with gene products of the major histocompatibility complex. Ann. Rev. Immunol. *3*, 237–261.

Schwartz, R. H., and Sredni, B. (1982). Alloreactivity of antigen-specific T-cell clones. In Isolation, Characterization and Utilization of Lymphocyte Clones, C. G. Fathman and F. Fitch, eds. (New York: Academic Press), pp. 375–384.

Sherman, L. A. (1982). Recognition of conformational determinants on H-2 by cytolytic T lymphocytes. Nature *297*, 511–513.

Sim, G. K., and Augustin, A. A. (1985). V_β gene polymorphism and a major polyclonal T cell receptor idiotype. Cell *42*, 89–92.

Sim, G. K., Yague, J., Nelson, J., Marrack, P., Palmer, E., Augustin, A., and Kappler, J. (1984). Primary structure of human T-cell receptor α-chain. Nature *312*, 771–775.

Singer, A., and Hodes, R. J. (1982). Major histocompatibility complex-restricted self-recognition in responses to trinitro-phenyl-ficoll. Adaptive differentiation and self-recognition by B cells. J. Exp. Med. *156*, 1415–1434.

Spits, H., Yssel, H., Voordouw, A., and de Vries, J. (1985). The role of T8 in the cytotoxic activity in cloned cytotoxic T lymphocytes lines specific for class II and class I major histocompatibility complex antigens. J. Immunol. *134*, 2294–2298.

Sprent, J., and Bruce, J. (1979). Lymphoid function in $F_1 \rightarrow$ parent chimeras. Lack of evidence for adaptive differentiation of B cells or antigen-presenting cells. J. Exp. Med. *150*, 715–720.

Sredni, B., and Schwartz, R. H. (1981). Antigen-specific, proliferating T lymphocyte clones: methodology, specificity, MHC restriction and alloreactivity. Immunol. Rev. *54*, 187–224.

Stockinger, R., Bartlett, K., Pfizenmaier, K., Röllinghoff, M., and Wagner, H. (1981). H-2 restriction as a consequence of intentional priming. J. Exp. Med. *153*, 1629–1639.

Streilein, J. W., and Duncan, W. R. (1983). On the anomalous nature of the major histocompatibility complex in Syrian hamsters, Hm-1. Transplant. Proceed. *15*, 1540–1545.

Stroynowski, I., Orn, A., Goodenow, R., McMillan, M., Forman, J., Brayton, P., Frelinger, J., and Hood, L. (1984). Cytotoxic T lymphocytes recognize determinants on the Balb/c-H-2L^d molecule controlled by α1 and α2 but not α3 external domains. Immunogenetics *20*, 141–154.

Sukhatme, U. P., Sizer, K. C., Vollmer, A. C., Hunkapiller, T., and Parnes, J. R. (1985). The T cell differentiation antigen Leu-2/T8 is homologous to immunoglobulin and T cell receptor variable regions. Cell *40*, 591–597.

Swain, S. L. (1981). Significance of Lyt phenotypes: Lyt2 antibodies block activities of T cells that recognize class I major histocompatibility complex antigens regardless of their function. Proc. Natl. Acad. Sci. USA *78*, 7101–7105.

Swain, S. L., Dutton, R. W., Schwab, R., and Yamamoto, J. (1983). Xenogeneic human anti-mouse T-cell responses are due to the activity of the same functional T-cell subsets responsible for allospecific and major histocompatibility complex-restricted responses. J. Exp. Med. *157*, 720–729.

Tonegawa, S. (1983). Somatic generation of antibody diversity. Nature *302*, 575–581.

Townsend, A. (1985). Molecules at work on the T-cell surface. Immunology Today *6*, 68–70.

van den Elsen, P., Shepley, B., Cho, M., and Terhorst, C. (1985). Isolation and characterization of a cDNA clone encoding the murine homologue of the human 20 K T3/T-cell receptor glycoprotein. Nature *314*, 542–544.

von Boehmer, H., Hengartner, H., Nabholz, M., Henhardt, W., Schreier, M., and Haas, W. (1979). Fine specificity of a continuously growing killer cell clone specific for H-Y antigen. Eur. J. Immunol. *9*, 592–597.

Wagner, H., Hardt, C., Bartlett, R., Stockinger, H., Röllinghoff, M., Rodt, H., and Pfizenmaier, K. (1981). Frequency analysis of cytotoxic T-lymphocyte precursors of chimeric mice. J. Exp. Med. *153*, 1517–1532.

Waldman, H., Pope, H., Brent, L., and Bighouse, K. (1978). The influence of major histocompatibility complex on lymphocyte interactions in antibody formation. Nature *274*, 166.

Wettstein, P. J., Bailey, D. W., Mokraaten, L. E., Klein, J., and Frelinger, J. A. (1978). T-lymphocyte response to H-2 mutants. J. Exp. Med. *147*, 1395–1404.

Winoto, A., Mjolsness, S., and Hood, L. (1985). Genomic organization of the genes encoding mouse T-cell receptor α-chain. Nature *316*, 832–836.

Wylie, D. E., Sherman, L. A., and Klinman, N. R. (1982). Participation of the major histocompatibility complex in antibody recognition of viral antigens expressed on infected cells. J. Exp. Med. *155*, 403–414.

Yagüe, J., White, J., Coleclough, C., Kappler, J., Palmer, E., and Marrack, P. (1985). The T cell receptor: the α and β chains define idiotype, and antigen and MHC specificity. Cell *42*, 81–87.

Zamoyska, R., Vollmer, A. C., Sizer, K. C., Liaw, C. W., and Parnes, J. R. (1985). Two Lyt-2 polypeptides arise from a single gene by alternative splicing patterns of mRNA. Cell *43*, 153–163.

Zinkernagel, R. M. (1982). Selection of restriction specificities of virus-specific cytotoxic T cells in the thymus: no evidence for a crucial role of antigen-presenting cells. J. Exp. Med. *156*, 1842–1847.

Zinkernagel, R. M., Callahan, G. N., Althage, A., Cooper, S., Klein, P. A., and Klein, J. (1978). Thymus in differentiation of "H-2 self-recognition" by T cells: evidence for dual recognition? J. Exp. Med. *144*, 933.

Zinkernagel, R. M., and Doherty, P. C. (1975). H-2 compatibility requirement for T-cell-mediated lysis of target cells infected with lymphocytic choriomeningitis virus: different cytotoxic T-cell specificities are associated with structures coded for in H-2K or H-2D. J. Exp. Med. *141*, 1427–1436.

V. DIVERSITY, EVOLUTION AND THE IMMUNOGLOBULIN GENE SUPERFAMILY

THE IMPACT OF MODERN GENETICS ON EVOLUTIONARY THEORY

Tim Hunkapiller, Henry Huang, Leroy Hood, and John H. Campbell

Evolution is a tangible and dynamic aspect of biological communities. However, while evidence of its occurrence is overwhelming, its precise nature is much more obscure. The pioneering work of Haldane, Fisher, Wright, and others in the 1920s and 1930s laid the theoretical framework for population genetics. From this foundation, Huxley, in the early 1940s, developed the first widely accepted model for evolutionary change and the forces involved—the "modern synthesis" or Neodarwinism. This theory is essentially a marriage of Mendelian genetics and classic Darwinian selection and has been a cornerstone of modern biological thought. Today, however, the theory of evolution as embodied by Neodarwinism is a theory in some turmoil. There are phyletic gradualists and punctuated equilibrists. There are strict selectionists and just as strict neutralists. There are altruistic genes that turn out to be selfish. There are even reports that acquired traits can be inherited. Unfortunately, classic paleontological and taxonomic approaches to these and other issues and problems have yielded results too equivocal for general consensus. Even the refinement of taxonomy to the biochemical level has stirred as much controversy as it has shed light.

In just the last few years, exciting technological advances in the field of molecular genetics have been responsible for a revolutionary new view of the structure and organization of genetic information.

Likewise, this new perspective has suggested many intriguing and novel means of genetic variation. It is our belief that this new model of the genome imposes both constraints and potentialities on any new synthesis of evolutionary theory and should aid in resolving many of the outstanding issues of debate. In this chapter we will present a brief outline of several of the phenomena responsible for this new genetic model, in order to convey something of the current view of the complex nature of genetic elements, their functional and structural organization within the genome, and the modes of physiological variation of germline information.

The genetic elements that we will discuss represent, by themselves and as components, different levels of organizational complexity that generate hierarchical strategies of function and fitness. Variation, it seems, is an integral part of the dynamic equilibrium between these systems. The rate and style of evolution then depends not only on the external forces exerted upon organisms but also upon the nature of variation prescribed by their genetic organization.

We will start our discussion with a brief description of some particular genomic structural elements and the implications they have for genetic variation. We will then describe examples of the higher level functional organization of these components and the possibility they provide for integrated genetic change.

SPLIT GENES—EXONS AND INTRONS

Perhaps no observation has come as such a great surprise to molecular biologists as the fact that eukaryotic genes are split into a series of alternating peptide-coding regions (exons) and intervening sequences (introns) that are transcribed together as a single, high molecular RNA species. Known exon numbers range from two to as many as 52 for α-collagen (for review, see Abelson, 1979). Subsequent enzymatic excision of the noncoding information from the RNA transcript is required to generate a functional messenger RNA in which all coding information is contiguous and translatable. This process is called RNA splicing (Figure 1).

The discovery of split genes was followed by the corollary observation that often functionally discrete domains of particular proteins are coded for by individual exons (Blake, 1981). The simplest and perhaps clearest example of this phenomenon is seen in the antibody system where discrete exons encode homologous structural domains of the antibody molecule (Calame et al., 1980; Sakano et al., 1979) (Fig-

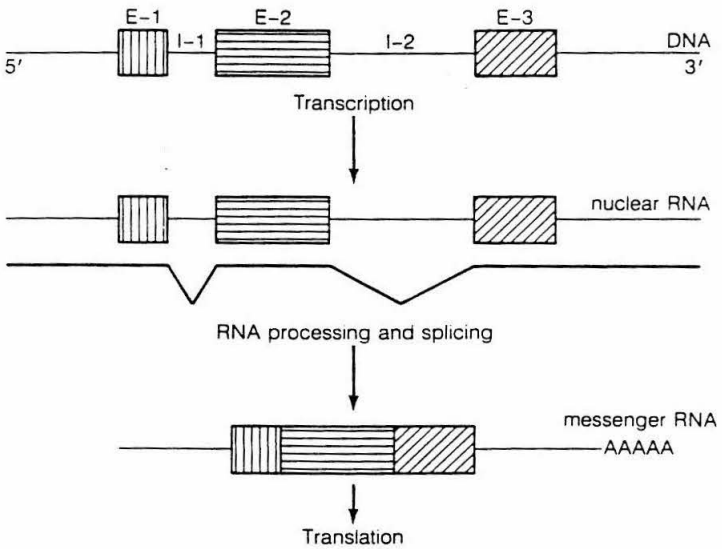


FIGURE 1. A model of RNA splicing. The original RNA transcript of a split gene includes both coding (exons or E) and noncoding (introns or I) sequences. RNA processing, which includes the splicing together of all exons and the subsequent loss of introns, produces a mature, translatable messenger RNA. Non-protein coding RNAs (transfer RNAs, ribosomal RNAs) undergo a similar RNA splicing process.

ure 2). In this context, the evolution of many multidomain proteins, like antibodies, can readily be explained by the duplication of a single primordial exon and the subsequent divergence of the homologs. Thus, exons encode evolutionary as well as functional units in proteins. The duplication of functional domains without exon splicing would require the far less likely event of precisely contiguous and in-phase alignment of the duplicate and original coding sequence. Even more striking are the evolutionary implications of the rearrangement of exons referred to as "exon shuffling" (Doolittle, 1978; Darnell, 1978; Gilbert, 1978; Crick, 1979). When genes are composed of multiple discrete exons that encode distinct functional or structural peptide domains, new combinations of exons can generate novel synergistic combinations of these domains. Any intra- or intergenic recombinational event or mutation in splice-site signal sequence can generate unique permutations of this kind. Isolation of the coding information as exons surrounded by nontranslated DNA sequence helps maintain the integrity of genetic information during such events and insures a proper reading frame alignment. This evolutionary strategy removes the requisite of reinventing the wheel or rather reevolving certain structural and/or func-

tional strategies before they are employed within significantly new contexts. For example, recombination of the exon encoding the membrane-anchoring domain of one gene with a second gene whose product previously was secreted drastically alters the functional context of the second protein. Exon shuffling would allow significant evolution of proteins to proceed in major steps rather than through a series of intermediates reflecting minor incremental changes. Since the domains involved in exon shuffling would have already experienced selection for their structural and functional viability, selection on new combinations will need to operate less on this aspect and more on the functional relationship of the new product and the cell producing it. This view does not imply that eukaryote genomes are organized to encourage such evolution, but rather that the evolution is a fortuitous by-product of an arrangement that serves other primary functions.

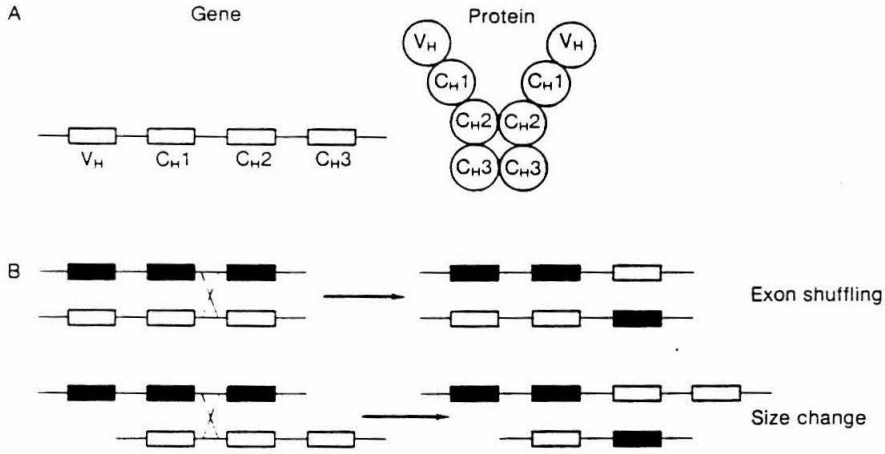


FIGURE 2. A. The correlation of immunoglobulin exons and globular domains. The DNA organization of an immunoglobulin heavy chain gene and domain structure of the corresponding heavy chain polypeptide are represented. The boxes represent exons, and the lines represent flanking sequences or introns for the gene. The circles represent discrete globular domains for the protein (the light chains are omitted for clarity). There is a perfect correlation between the exons and the corresponding globular domains. B. Exon shuffling. Classical genetic phenomena can result in the rearrangement or "shuffling" of exon information. Shown are possible rearrangement events involving immunoglobulin heavy chain constant region genes that could result in new exon arrangements and, subsequently, new synergistic associations of functionally and/or structurally discrete domains.

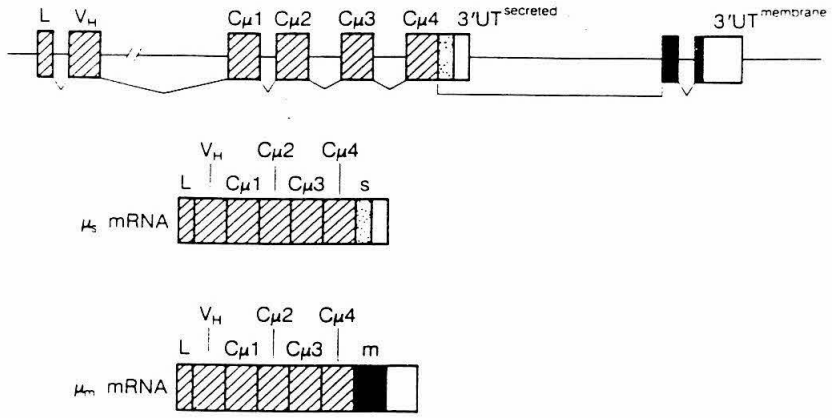


FIGURE 3. Antibody molecules are expressed in either the secreted (s) or the membrane-bound (m) form, depending on the RNA splicing and processing pathways employed after transcription of the genes. The exons (boxes) and introns (lines) of the μ heavy chain gene (encoding antibodies of the IgM class) are depicted. Alternative splicing patterns result in two mRNAs, $\mu_{membrane}$ and $\mu_{secreted}$, that differ only at the 3′ (C-terminal) coding end. Open boxes denote the 3′ untranslated regions. Black boxes represent the membrane tail coding region and stippled boxes the secreted tail-coding region. L, V_H, and C_μ denote the leader-, variable-gene-, and constant-domain exons, respectively.

Not only can new exon permutations be generated, but developmental expansion of the information content of any given germline sequence is possible through the use of alternative RNA splicing pathways. As shown in Figure 3, antibody heavy chain genes have distinct coding regions for two relatively small, alternative carboxy-terminal tails. One of these coding regions (T_S) encodes a peptide sequence associated with the secreted form of the antibody molecule that carries out the effector functions of humoral immunity. The second of these coding regions (T_M) encodes a membrane tail which inserts into the lipid bilayer of the cell membrane and renders the antibody molecule a cell-surface receptor for triggering subsequent steps of differentiation upon interaction with a foreign molecular pattern (antigen). Though both exons are arrayed sequentially with other heavy chain exons, two alternative modes of RNA splicing ensure that only one of the two tail-exons is included in a mature message RNA (Kehry et al., 1980; Early et al., 1980; Rogers et al., 1980).

The exon-intron arrangement is complicated even further. One gene's intron may in part code for a different, structurally unrelated protein. Specifically, it has recently been shown that a protein involved in the processing of cytochrome *b* mRNA in yeast ("mRNA maturase")

actually resides within the intron of the gene it helps regulate (Lazowska et al., 1980).

The hypothesis of evolutionary exon shuffling readily permits us to understand how extremely sophisticated gene systems can arise through relatively few genetic events. Therefore, under the appropriate environmental conditions, such systems also can arise in relatively short periods of evolutionary time. Let us now consider an analogous but higher order of organization for eukaryotic information that is characteristic of virtually all eukaryotic systems that have been studied to date, the multigene family.

MULTIGENE FAMILIES

Many eukaryotic genes are found in groups or families of multiple, related copies. These multigene families are made up of information units that (1) are homologous in structure, (2) overlap in function, and (3) are generally tandemly linked (Figure 4) (for review, see Hood et al., 1975). Multigene families range in size from the few gene copies of the hemoglobin families to thousands of copies of ribosomal genes and satellite DNAs. The gene copies within a family can be virtually identical or can vary markedly from one another. Multigene families are evolutionarily dynamic entities that provide special opportunities for processing, expression, expanding, and evolving information. Certain general properties of multigene families have important evolutionary implications.

1. *Variation in gene copy number.* Multigene families often vary in size from one species to another. For example, the clawed toad *Xenopus laevis* has 24,000 copies of 5S ribosomal genes, whereas its counterpart *Xenopus borealis* has 6,000 copies of the same information (Brown et al., 1971). Presumably the expansion and contraction of multigene families is promoted through homologous but unequal crossing-over (Figure 5) and can occur relatively rapidly. Variation in copy number on this scale has important implications as to the variation between copies and will be discussed shortly.

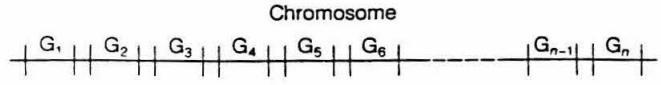


FIGURE 4. Model of a multigene family. (From Hood et al., 1975.)

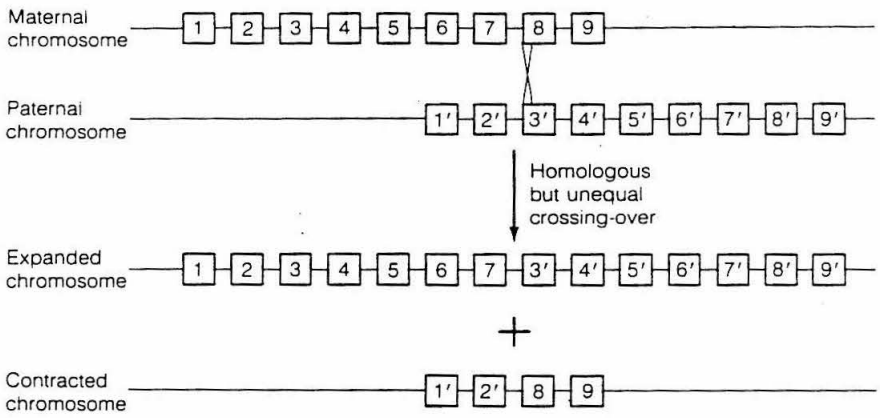


FIGURE 5. A model of homologous but unequal crossing-over. Homologous crossing-over events can be promoted by the mispairing of homologous genes between chromosomes carrying closely linked, homologous multigene families. This process yields one chromosome with an increased number of genes and a second with a decreased number of genes. (From Hood et al., 1975.)

2. *Heterogeneity.* A single gene can generally encode only a single, chemically homogeneous gene product. In contrast, a multigene family can encode a heterogeneous collection of closely related protein species. For example, the hemoglobins synthesized by humans are encoded by two multigene families (Figure 6A), the α family and the β family, whose individual genes are expressed in a developmentally regulated fashion (Figure 6B) (Bernards et al., 1979; Efstratiadis et al., 1980; Weatherall and Clegg, 1979; Proudfoot et al., 1980). Accordingly, distinct globins are expressed at the embryonic ($\zeta_2\epsilon_2$, $\alpha_2\epsilon_2$), the fetal ($\alpha_2\gamma_2$), and the adult ($\alpha_2\delta_2$, $\alpha_2\beta_2$) stages of development. Interestingly, the developmental order of β-like gene expression directly reflects the β family genomic order (5′ to 3′). Also, different globins have distinct oxygen-binding properties and are regulated by various effector molecules in quite different fashions. Thus, microheterogeneity has evolved which is physiologically useful and is expressed in a precisely regulated fashion. Certainly the most extreme example of heterogeneity in the gene families analyzed to date is the diversity seen among the antibody genes, which provides the basis for the enormous range of specificities in the immune system.

3. *Multigene families shield their gene copies from natural selection.* Gene copies that overlap in function are shielded from selection, in the sense that deleterious mutations of one gene copy will be buffered if other gene copies can assume at least in part the mutated

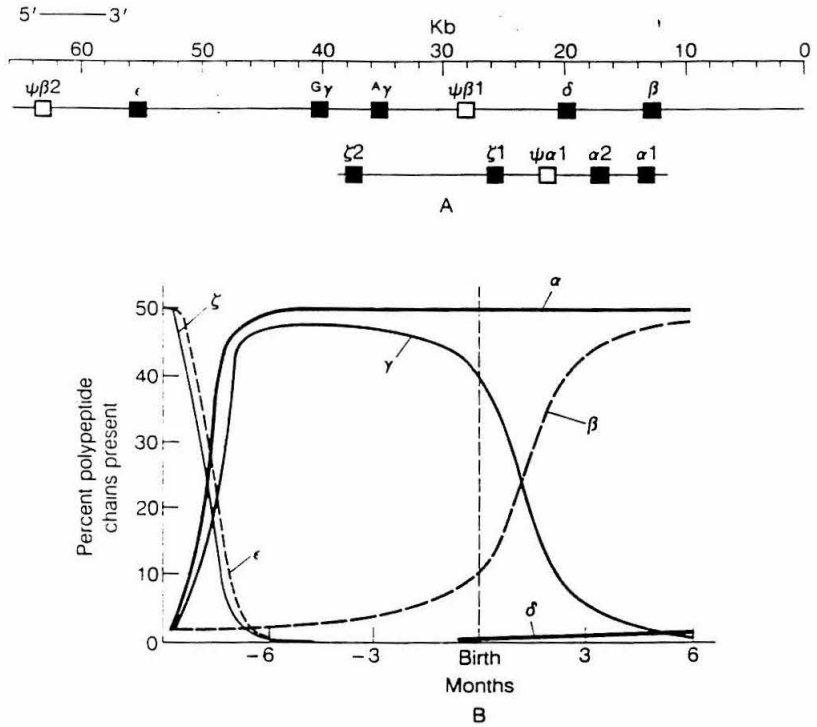


FIGURE 6. A. Models of the beta (β) and alpha (α) globin gene families in man. Individual genes (boxes) denote separate genes and do not indicate exon-intron configurations. Open boxes represent pseudogenes (ψ). The ϵ and ζ genes are embryonic; the γ genes are fetal; the α, β, and δ genes are adult. The size of these families is indicated in kilobases (kb). (Courtesy of T. Maniatis.) B. Time course of concentrations of the respective polypeptides.

gene's function. For example, hundreds of virtually identical histone genes are present in most higher organisms. The effect on the organism from the loss of function of one histone gene copy through mutation would be negligible. Thus, redundant gene copies would be expected to diverge spontaneously over evolutionary time. Besides promoting this divergence, shielding has at least two other interesting consequences. First, pseudogenes arise and appear to be a characteristic feature of multigene families (see Figure 6A). A pseudogene is one whose sequence has been altered so that it no longer is functionally expressed at the polypeptide level. Pseudogenes arise through mutations generating stop codons within exons, changes in the RNA pro-

cessing signals, or sequence interruptions which lead to changes in the reading frame. Indeed, 25 percent of known globin genes appear to fall in this category. The frequency of pseudogenes underscores the fact that individual members of a gene family may be under reduced selective restraint. Presumably pseudogene sequences will eventually randomize through mutation unless they have selectable functions (e.g., regulation) within the gene family (Li et al., 1981). Second, and most important, selection will tend to operate on the function of the multigene family taken as a whole rather than operating on the function of the individual gene copies that comprise that family. Thus, in one sense, the unit of selection becomes the multigene family rather than a single Mendelian gene.

4. *Coincidental evolution.* Homologous multigene families in distinct evolutionary lines can evolve in a coincidental fashion. Coincidental evolution denotes the tendency for genes within the same gene family to evolve in a similar or coincidental fashion (Figure 7). Coincidental evolution may arise either by homologous but unequal crossing-over or some type of gene conversion (Figure 8). Thus when genes are organized into tandemly linked families, there are mechanisms (Figure 8) which often insure coincidental evolution and, accordingly, limit variation.

The inherent tendency of gene copies in a multigene family to evolve coincidentally opposes their reciprocal tendency to diverge, due to their shielding from natural selection. As has been noted previously, some gene families containing hundreds of members, such as the histones, are highly uniform in structure while others, especially the antibody genes, are remarkably heterogeneous. Explanations for differences in the heterogeneity may relate to the structural characteristics of particular gene families which might increase or decrease unequal crossing-over or gene correction and the propensity for coin-

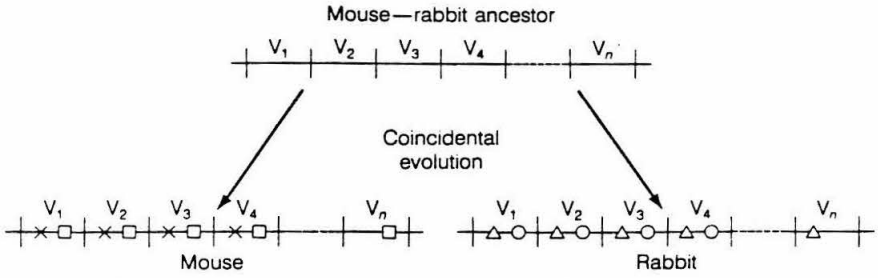


FIGURE 7. A diagram of coincidental evolution during the divergence of the mouse and rabbit evolutionary lines. ×, □, ○ and △ represent coincidental changes in the respective evolutionary lines. (From Hood et al., 1975.)

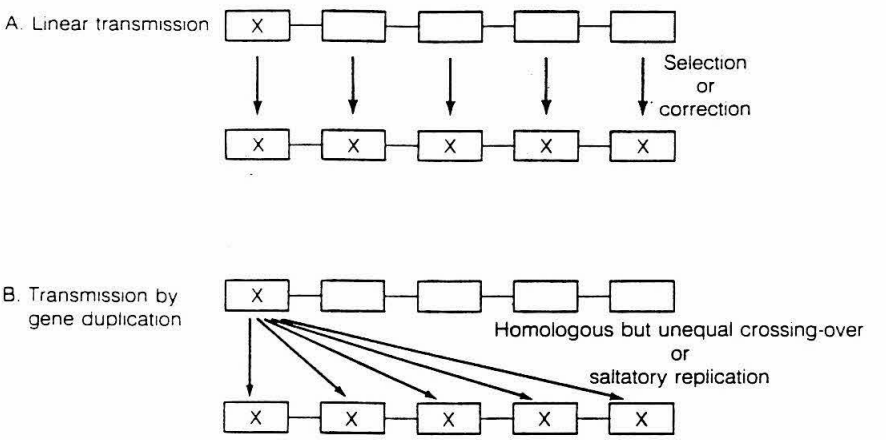


FIGURE 8. Two models of coincidental evolution. A. Linear transmission implies that all genes have been altered in a similar manner. B. Transmission by gene duplication, as in homologous but unequal crossing-over, indicates that particular gene copies have been expanded while others have been lost. (From Hood et al., 1975.)

cidental evolution. Alternatively, they may reflect structural and/or quantitative constraints at the gene product level. Copy homogeneity in gene families can be directly selected for when many functional gene copies are needed to express large quantities of product or when homogeneity of the gene product is critical for effective function or control. Ultimately the degree of homogeneity must reflect the balance between the constraints on variation imposed by the physical nature of the family and the information content necessary for that family to function.

5. *Information expansion by combinatorial and mutational mechanisms.* Multigene families can have informational and/or quantitative functions. The high copy number of ribosomal RNA genes, for example, insures the sufficient quantity of a very important, virtually homogeneous product. On the other hand, informational families like the globins produce heterogeneous products that allow an organism to fine tune its phenotype to various physiological signals (see Figure 6A, B). The antibody gene families code for specific antibody molecules against an amazingly broad spectrum of antigens. These families are predominantly informational systems that greatly expand their infor-

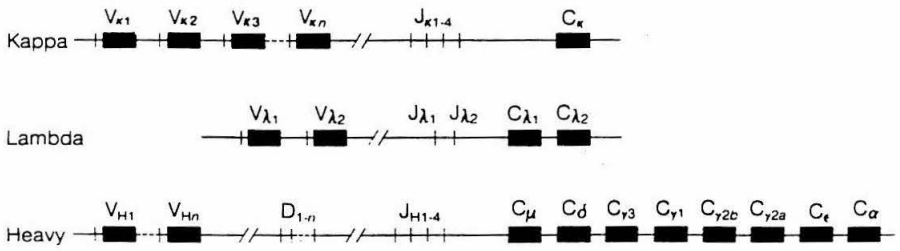


FIGURE 9. Organization of the three antibody gene families of the mouse. Exon-intron organization of heavy chain constant region genes is not shown. V, D, J, and C denote variable, diversity, joining, and constant genes, respectively. (From Early and Hood, 1981.)

mation content by DNA rearrangements and somatic mutations during development.

Antibodies are composed of two polypeptides, light and heavy chains, which in turn are divided into two distinct regions, an amino-terminal variable (V) and a carboxy-terminal constant (C) region. These polypeptides fold into four to five discrete domains which carry out distinct functions—pattern recognition by the variable domains and effector functions by the constant domains. The antibody molecules are encoded by three multigene families, two for light chains λ and κ, and a third for heavy chains (Figure 9). Light chains are

TABLE 1. Estimated numbers of antibody gene segments in the mouse and the diversity they can express via combinatorial mechanisms operating at the DNA and protein levels (see text).

1. GERMLINE		
	Kappa	~250 V_κ
		4 J_κ
	Heavy	~250 V_H
		~10 D
		4 J_H
	Lambda	2–3 V_λ
		2–3 J_λ
2. COMBINATORIAL JOINING		
	Kappa	250 $V_\kappa \times 4\ J_\kappa$ = 1000 κ genes
	Heavy	250 $V_H \times 10$ D $\times 4\ J_H$ = 10,000 H genes
3. COMBINATORIAL ASSOCIATION		
	1000 κ × 10,000 H = 10^7 antibody molecules	

encoded by three separate gene elements, variable (V_L), joining (J_L), and constant (C_L). Heavy-chain gene families include a fourth segment, diversity (D). Table 1 gives our current estimates as to the number of germline gene segments within each of the three antibody gene families of mice.

The initial expression of antibody genes requires DNA rearrangement events that bring into contiguous apposition the elements of V_L and V_H genes with the subsequent loss of any intervening DNA sequence (Figure 10) (Brack et al., 1978; Seidman et al., 1978; Davis et al., 1980*a*). In light chains V_L and J_L gene elements are joined, while in the heavy chains, the V_H, D, and J_H elements are joined together. These rearrangements are presumed to occur during the development of particular antibody-producing (B) cells. Table 1 shows that unrestricted combinatorial joining of the gene elements within the antibody gene families would be equivalent to having 1,000 V_κ and 10,000 V_H germline genes in one organism. In the heavy-chain family, the joined V_H gene can be juxtaposed with different C_H segments through a developmentally regulated DNA rearrangement called class switching (Figure 10). In this way, identical antigen specificity can be associated with the various effector functions encoded by the different C_H sequences. The combinatorial association of the light and heavy

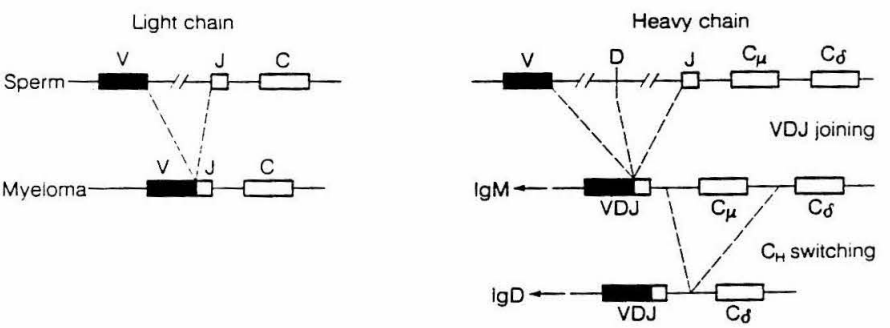


FIGURE 10. Two types of DNA rearrangements which occur during the differentiation of antibody-producing cells. In light and heavy chain genes V-J or V-D-J joining juxtaposes the gene segments encoding the V_L and V_H genes. Subsequently in heavy chain genes class, or C_H, switching may occur. The class switch leads to the expression of different immunoglobulin classes. Sperm indicates DNA undifferentiated with regard to antibody function, whereas myeloma denotes DNA in which one or more DNA rearrangements have occurred.

polypeptide chains is necessary to generate functional antibodies. Thus, the combinatorial pairing of light and heavy chains, at the protein level, also permits an enormous expansion of the information potential of antibody gene families and, presumably, the functional range of antigen-binding specificities. Indeed, from Table 1 we can see that 10^7 distinct antibody molecules could be generated through combinatorial mechanisms operating at both the nucleic acid and the protein level.

The mechanism that joins together the distinct gene elements has another, more subtle consequence for antibody diversity. The site of joining is flexible and may be at many different points in the junctional sequences—thereby leading to hybrid codons and junctional regions of varying size (Weigert et al., 1980; Sakano et al., 1981). Junctional diversity occurs in an area of the light and heavy chains that has a particularly important contribution to the antigen-binding site. A final, less well-understood form of antibody diversification is a mutational mechanism that appears to operate only during a narrow time span of B-cell differentiation and is probably activated by antigen stimulation (Gearhart et al., 1981; Crews et al., 1981). It causes the equivalent of point mutations in the rearranged variable gene elements and correlates with heavy-chain class switching. This mutational mechanism expands enormously the variability and hence the functional information of antibody genes and their corresponding polypeptide products.

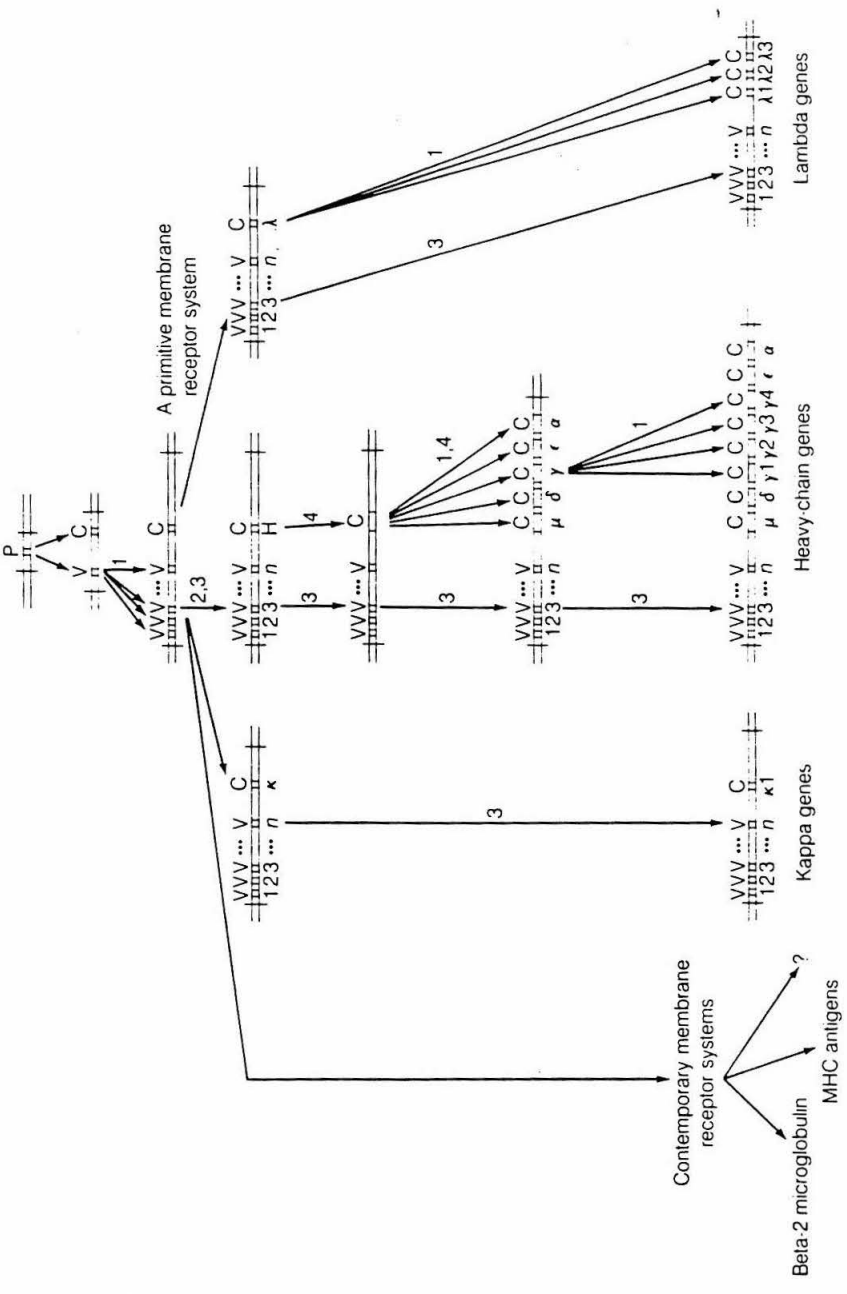
Although presently demonstrated only in antibody gene families, these combinatorial and mutational mechanisms for information expansion may well be employed in other complex eukaryotic systems, such as those seen in the nervous system. In this regard, two points are striking. First, the antibody system employs strategies that operate in somatic cells to vastly increase the amount of useful information carried by the germline, but which are directly dependent on the physical nature of the germline families. Hence, gene structure and organization that allows such information amplification is directly selectable. Second, the existence of enzyme systems for altering DNA in these somatic cells poses the intriguing question as to whether similar mechanisms also exist for changing genetic information in germ cells.

6. *Multigene families are a new unit of evolution.* The emergence of informationally complex multigene families may have paralleled that of metazoan organisms (Figure 11). Indeed, these multigene families may reflect the enormous increase of information required at the metazoan cell surface for cellular interactions, transmembrane signaling, and cellular migrations. Once generated, these families presented the organism with a new evolutionary unit—the multigene

family itself. A multigene family may be duplicated in part or *in toto*. Duplicate gene families could then diverge just as duplicate genes do and come to encode new and complex aspects of eukaryotic phenotype. Since the duplication of a multigene family would include the attendant control mechanisms as well as structural genes, very sophisticated multigene families could arise in relatively short periods of evolutionary time, without the need to recreate the exhaustive and meticulous evolutionary process that was used in initially shaping the gene family. Thus, in time, a superfamily composed of many multigene families, each devoted to encoding distinct aspects of a eukaryotic phenotype, could emerge (Figure 11). This attractive hypothesis has recently been supported by the intriguing observation that genes encoding the transplantation antigens of mammals appear to be homologous to those of the antibody gene families (Strominger et al., 1980; Steinmetz et al., 1981). Thus, two complex mutigene families encoding various aspects of phenotype relating to development of the immune response appear to have a common evolutionary heritage. With the duplication of trans-active regulatory strategies, functional interaction and informational feedback schemes also can be maintained between disparate families. Again, these mechanisms permit very sophisticated gene families to evolve through a limited number of genetic events and possibly within a very short evolutionary time.

DYNAMIC GENOME—MOBILE GENETIC ELEMENTS

Perhaps the most difficult group of genetic phenomena to incorporate into the current evolutionary synthesis is represented by a staggering array of *quasi-stable genetic elements*. These are sequence units that are able to rapidly alter the genetic structure and information of an organism through rearrangement and/or expansion of genetic information, often in physiologic response to that organism's environment. Particularly interesting are mobile forms that are able to transfer genetic information between organisms, and even between species. Also, their frequent capacity for independent replication gives many of these elements a semiautonomous nature and makes them the purest example of selfish DNA: that is, DNA sequences within a cell that experience selective forces distinct from those experienced by the cell itself. Since the fitness of selfish DNA is to some degree independent of the fitness of the cell, selfish DNA is capable of independent evolution. It can be argued that all DNA possesses some selfishness. However, elements capable of interorganismic transmission are certainly less constrained by the selective needs of a particular host.


P
V
C
1
2,3
A primitive membrane receptor system
3
4
1,4
Lambda genes
Heavy-chain genes
Kappa genes
Contemporary membrane receptor systems
Beta-2 microglobulin
MHC antigens
?

◀ **FIGURE 11.** A hypothetical scheme for the evolution of certain informational multigene families. The order of gene duplication events is unknown. A number of genetic mechanisms seem to be employed in the evolution of these families, as indicated by numbers adjacent to arrows. These are (1) discrete gene duplication, (2) gene duplication by polyploidization or chromosomal translocation, (3) contiguous gene duplication, and (4) coincidental evolution of multiple genes. Mechanisms 1 and 4 may be identical. (Adapted from Gally and Edelman, 1972.)

The mobile quasi-stable systems range in size and complexity from the tiny viroids of plants and insertion (IS) sequences of bacteria to sophisticated viruses. Their effects upon the host organism range from altering the expression of single genes to subverting the entire genome to their own purposes. Although most of the well-studied examples of mobile quasi-stable elements are found in prokaryotes, analogous systems are known or suspected in eukaryotes as well.

The best known and most widely studied group of mobile genetic elements are the transposable elements of prokaryotes (for review, see Kleckner, 1977; Bukhari et al., 1977). They are relatively short segments of DNA with direct or inverted repeats at either end. They are characterized by their participation in "illegitimate recombination" events (independent of the cell's normal recombination machinery), which may involve the duplication and translocation of their genetic information. Some are highly site specific in their transposition, while others insert randomly throughout the host's DNA. Deletion occurs independently of transposition. Similar elements that are near one another can transpose as one unit and duplicate genetic information in the DNA sequence between them, thus generating larger and more complex assemblies of mobile elements. Therefore, transposable elements represent a spectrum of forms of increasing complexity. Within this spectrum, the simpler types occur both individually and as building blocks for more complex elements, and so on to even more complex units (Figure 12).

The simplest transposable elements in bacteria are insertion sequences (IS) about 1,000 base pairs (1 kb) long with direct repeats of about 150 base pairs at either end (Figure 12). Many different types of IS are known in *E. coli*, strains of which may harbor 6 to 10 copies of an IS element. A duplication event that inserts an IS element into or adjacent to a gene can positively or negatively attenuate or totally block the expression of the gene.

When a phenotypically recognizable gene becomes associated with a transposable element, the unit is called a transposon (Tn) (Figure 12). Transposons carry a wide variety of genes. Most often these genes are nonessential for the normal function of the cell, but they may play a critical role under conditions of stress. The best examples, and cause of much current medical concern, are transposons that carry genes which confer resistance to antibiotics.

Transposons can be quite simple. However, by fortuitously capturing genes involved in their own replication and expression, more complex entities can arise with greater selfish characteristics. For example, element Tn117 encodes an erythromycin-resistance gene that is under repressor control of another gene specified by the same transposon. This repressor also seems to regulate transposition. Thus, exposure to erythromycin not only derepresses the resistance gene that protects the host organism but also facilitates duplication of the element.

Plasmids are autonomously replicating molecules (10^1–10^2 kb) that exist as nonintegrated or extrachromosomal bodies within the host cell (Figure 12). Conjugative plasmids constitute still higher levels of organization. The first of these to be widely studied was the fertility element (F) of *E. coli*. This element is able to initiate conjugation, a

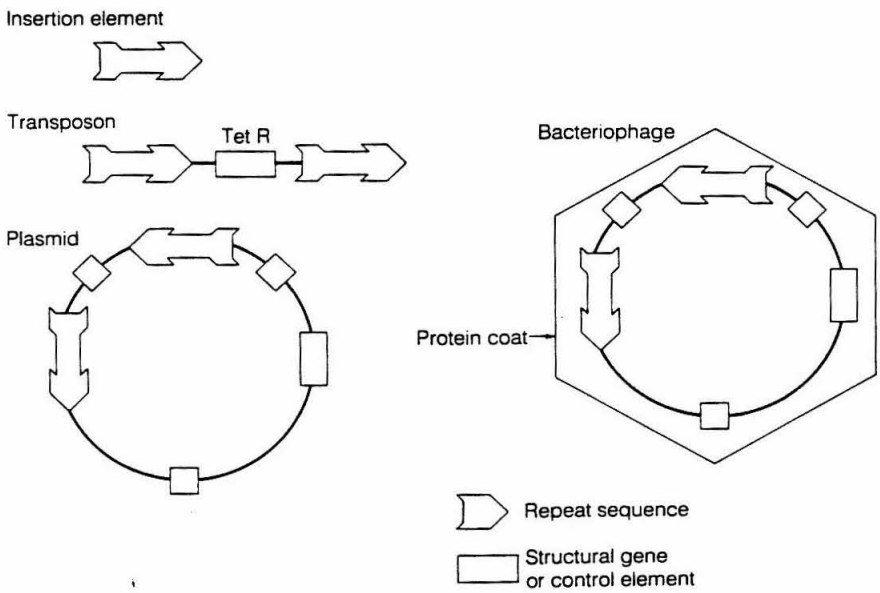


FIGURE 12. Four classes of prokaryote mobile genetic elements. Note the overlapping structural organizations.

process that directly transfers plasmic and chromosomal information of an F^+ cell into a recipient F^- cell. More promiscuous conjugative plasmids can even move between different species. Nonconjugative plasmids can be transferred to recipient cells during conjugation mediated by conjugative plasmids. Plasmids serve as vehicles for the cell-cell transmission of IS and transposons (Figure 12). Transposition from plasmid to host DNA can result in the stable transfer of phenotypes between the chromosomes of different organisms. Therefore, environmental conditions which promote transposition in one cell can directly facilitate the rapid acquisition of any transposition-linked trait by a very large cell population with little or no regard for that trait's selective value within the population.

The transmission of plasmids is a replication process in which the donor cell retains the same information transmitted to the recipient. Like transposons, plasmids tend to accumulate information that directly affects their own reproduction and fitness as well as the fitness of the host cells. Also, examples of both plasmids and transposons seem to have evolved through this cumulative process into still more complex units known as bacteriophages. Some members of this group are clearly plasmids (PI) or transposons (mu) that have acquired information necessary for a phage-type packaging, replication, and infection, while retaining the ability to behave in a plasmid or transposon fashion within the cell. These alternative modes of existence are determined by environmental conditions.

Many advanced phage like lambda are actually a mosaic of distinct chromosomal segments that display various modes of plasmid- and transposon-like replication, maintenance, and control. Nonlytic phage or those that integrate into the host genome also are able to provide phenotypic advantages to the cell (primarily in the form of resistance to further infection). Most important, phages represent extremely powerful vectors of genetic information between widely divergent cells.

Besides duplication of sequence information and disruption of resident gene activity, transposition is frequently associated with structural changes of the genomic material itself. Primarily, these are seen as deletions, duplications, and inversions at or near the site of transposition and may or may not include the mobile element (for review, see Nevers and Saedler, 1977).

In contrast to the above, some heritable recombination events are site specific and are directly involved in gene regulation. Inversion of a region of bacteriophage mu called the G loop appears to correlate with the production of infectious particles. In *Salmonella*, two genes (H-1 and H-2) code for the flagellar protein, flagellin. Phase variation

occurs when expression of one gene is replaced by expression of the other. Alternate expression of the H-1 and H-2 genes is controlled by an inversion event that changes the association between the H-2 gene and its promoter. In one orientation, H-2 is expressed along with a linked gene that encodes a repressor of H-1 expression. In the other orientation, only H-1 is expressed (Zieg et al., 1978). Interesting sequence homologies between this 1,000 bp element and Tn3 and bacteriophage mu suggest that phase variation might represent the evolutionary integration of a transposon-like selfish element into the physiologic economy of its host (Simon et al., 1980). A similar evolutionary relationship between transposable elements and many other prokaryote and eukaryote quasi-stable systems, including antibody rearrangements, has been suggested.

Although their overall importance is less well-understood, phenomena analogous to those just described for prokaryotes are displayed by eukaryotes as well as complex and poorly understood elaborations on these themes. Moreover, such phenomena appear quite common. Therefore, we believe quasi-stable elements also will play a significant evolutionary role in eukaryotic genomes.

The simplest quasi-stable eukaryotic systems are similar to bacterial insertion sequences and are exemplified by several dispersed repeated gene families in *Drosophila* (Copia, 412, etc.) (Finnegan et al., 1977; Potter et al., 1979). These elements are able to transpose with no clear pattern of distribution. They are maintained at about 25–35 copies per genome, with some variation for family type, and among individual cells, populations, and species. They are short (1 to 5 kb) with terminally redundant repeats (0.2 to 0.5 kb). As is the case with IS elements, they are frequently correlated with large chromosomal disruptions, such as deletions and inversions. Likewise, the elements may remain in place and initiate further local chromosomal disruptions. The control of transposition of these elements is not understood.

Although these elements generate no obvious protein product, members of these families are highly represented in the cytoplasmic pool of poly(A)$^+$ RNA. Consequently, some investigators have suggested a regulatory function for them. As a class, they make up perhaps 1 percent of the *Drosophila* genome. Similar DNA sequences and quasi-stable genetic elements are seen in yeast (Cameron et al., 1979) and appear to be present in many other eukaryotes. These and similar phenomena in eukaryotes are referred to specifically as mobile elements.

Eukaryotic mobile elements also exhibit a wide range of complexities analogous to prokaryote elements. For example, a series of mutations in the *white* locus (eye color) of *Drosophila* seem to be mediated by a transposon-like element derived from Copia (Rasmuson et al.,

1974). However, their nature and diversity do not always suggest the simple evolutionary connection obvious between many prokaryote mobile elements. Eukaryotes seem, in fact, to employ an even more diverse repertoire of mobile strategies.

The first phenomena to be suspected of a quasi-stable nature in any type of organism were the controlling elements of maize first studied by Rhoades and McClintock and others more than thirty years ago (for review, see Nevers and Saedler, 1977). Their apparent movement about the maize genome affected the regulation and expression of a wide variety of genetic markers. Furthermore, these changes were inherited. There are two separate component elements of maize controlling elements: (1) a receptor that inhibits gene expression and (2) a regulator that controls, transactively, the excision and transposition of itself, and the receptor. These maize-controlling elements can exist independently or linked together into an "autonomous" unit that can perform both functions. Note that the regulator has the ability to affect its own copy number in a heritable way.

In contrast to the apparently chaotic genetic changes associated with certain eukaryotic mobile elements, other quasi-stable genetic changes may be precisely controlled for gene regulation. For example, *Saccharomyces cerevisiae* (yeast) haploid cells each express one of two distinct mating types. Clonal derivatives of one type can switch to express the alternate form (for review, see Strathern and Herskowitz, 1979). Switching occurs in a cyclical manner and is under control of many genes. Mating type switch is the result of the physical replacement of one *cassette* of sequence information at the single mating-type expression locus with a new cassette from either of two alternate silent loci (Hicks et al., 1979). The silent master cassettes are not affected by mating type switching. The new mating type is then heritable.

Trypanosomes exhibit a similar cassette system for serotype determinants (Pays et al., 1981). Offspring of a single clone are able to express any one of an array of variant protein antigens by replacing the previously expressed sequence with a copy of a different one. This serotype is then inherited as a quasi-stable trait of that lineage. The original copy remains intact. The ability to vary the protein is important to the trypanosome's ability to escape the immune system of its host. The significance of all these examples is that the eukaryote genome can possess heritable traits that affect the nature of its own variation.

All examples of eukaryote quasi-stable genetic changes discussed so far produce changes that are, presumably, only vertically transmittable. The eukaryotic genome is incredibly complex, and the

metazoan nature of most studied eukaryotes makes horizontal transmission of information between germlines very difficult to study. Certainly, plasmid analogs are difficult to find, except perhaps in yeast. However, animal and plant viruses are abundant and could be quite significant in the horizontal transmission, both within and between organisms of information that could then be inherited. Integration of viral information within eukaryote genomes is well documented. In fact, 0.1 percent of the murine genome is thought to be related to the sequence of retroviruses, a class of RNA tumor viruses. Related viral sequences have been found in most vertebrate genomes studied and are closely identified with highly conserved repeat elements that resemble those of transposons (for review, see Bishop, 1978). Presumably, the repeat elements facilitate the integration of the viral sequence into the host genome and perhaps even transposition. Many human ailments are thought to result from the activation of information from these and other types of viruses that is normally present but latent in the genome.

Two other examples of horizontal transmission are particularly intriguing, because they imply exchange of genetic information over an incredibly wide phylogenetic range. *Agrobacterium tumefaciens* is a gram-negative soil bacterium that is the agent of crown gall neoplasia in many dicotyledonous plants. Oncogenic varieties of *A. tumefaciens* possess a large plasmid (Ti) that is able to conjugatively transfer to the cells of infected plants. Much of the information is then integrated into the host cell genome where it is ready to direct the synthesis of bacteria-related products (Chilton et al., 1976; Zambryski et al., 1980). Another class of genetic disease normally associated with plants involves the transfer of tiny agents called viroids. These are small (1 to 10 kb), naked single-stranded RNAs that can somehow subvert cellular functions upon infection. Although they have been recognized in plants for some time, recently certain animal diseases also have been linked to viroid-like agents. Very little is known about their biology or origin, however. An interesting hypothesis concerning their nature is that one organism's viroid may be another's discarded RNA. It remains to be seen whether or not viroids or bacterial plasmids can transfer heritable information into eukaryote genomes. Ultimately, however, there is no reason to believe eukaryote genomes are less prone to violation than prokaryote genomes.

The existence and nature of quasi-stable genetic changes have three particularly important theoretical implications. First, mobile elements can attain the status of a parasite of the genome. This is particularly so if originally, by its physical nature, the mobile element mimics necessary genetic elements and becomes a competitive substrate for the enzymatic and replicative mechanisms of the host. If transposition of an element is frequent enough, a population can be

parasitized with a genetic element that is irrelevant or even deleterious to the host. Also, heritable mutations within an element that affects its transposability can bring to bear unique, intragenic selective pressures and initiate competition between various transposable elements with little regard to the phenotype of the host. Of course, as with most symbiotic relationships, the parasite-host interaction will tend toward a mutualistic relationship over evolutionary time, as any benefit that the element can provide its host will reflect upon its own fitness. Thus, transposable elements are expected to evolve roles in the economy of the host genome even if they also have the capacity to evolve toward their own selfish ends. Second, transmissible elements can be seen as mobile extensions of the genome that maintain direct genetic continuity between divergent organisms. By lateral transmission, they can promote the rapid spread of selectively advantageous (e.g., antibiotic resistance) or even deleterious traits throughout populations and between species. Third, these types of recombination can generate novel genic and phenotypic arrangements with a single genetic event. These new arrangements can affect regulatory mechanisms for existing traits or even introduce entire new gene products.

Taken together, these phenomena describe a remarkably dynamic genome, one that can respond to its environment and affect its own variation. Genetic heritage must be seen as a compromise between the various selfish activities of the individual components and the selective needs of the whole.

HIERARCHICAL ORGANIZATION OF GENE EXPRESSION

The genome is dynamic but not chaotic. It is an interactive and fluid informational system best represented by hierarchical relationships. This hierarchical organization of genetic systems results in large sets of information being under common and thereby coordinate regulatory control. Therefore, it is possible for even one or very few genetic loci to control the expression of large numbers of other genetic loci. This is best exemplified by the *bithorax* complex. The *bithorax* complex is a cluster of at least 12 genetic loci that control the developmental fate of the majority of the body segments of *Drosophila melanogaster* (Lewis, 1981). *Drosophila* is a dipteran insect, characterized by having two wings on the second or mesothoracic segment and a pair of halteres on the third or metathoracic segment. The halteres are homologous to the pair of wings on the metathoracic segment on most other insects. Flies homozygous for bx^3 and *pbx*, mutant alleles at two loci in the *bithorax* complex, have four wings and no halteres (Lewis, 1963). Thus,

two genetic changes in the *bithorax* complex are sufficient to effect an enormous phenotypic change, presumably through the suppression of "haltere genes" and the activation of a large number of genetic loci involved in the construction of wings. Similar transformation of other segments through genetic changes in the *bithorax* complex are also seen, resulting, for example, in eight-legged flies (Lewis, 1981).

The hierarchical organization of gene expression during development has important implications for evolutionary theory. First, it makes possible major phenotypic changes that result from relatively few genetic changes, and thus rapid evolution of morphology can readily occur, provided selective conditions are favorable or neutral. It is probable that during the evolution of the Insecta, genetic changes in regulatory genes, perhaps only very few in number, created two-winged forms from the primitive four-winged state, and thus gave rise to the dipterans. It is certainly attractive to view even large-scale phyletic experiments like expansion or loss of segmentation in this context.

Second, although the new pair of wings in homozygous $bx^3\,pbx$ flies are smaller than normal (Lewis, 1951, 1963), they are remarkably faithful copies of the regular wings [$bx^3\,pbx$ plus mutation at a third locus, *abx,* renders the new pairs of wings virtually identical to the normal pair (Lewis, 1981)]. This leads us to believe that the second pair of wings are not the result of the activation of a cryptic set of "wing genes" specific for the metathoracic segment. Such cryptic genes should have been lost some time during the long evolutionary history of dipterans. Instead, we believe that the same set of "wing genes" are activated in a coordinated fashion in both the meso- and metathoracic segments during the development of the mutant flies. Thus the potential for the change from cryptic to manifest can be maintained for long periods of time if the genes involved are expressed elsewhere in the organism, and therefore are maintained by selection.

This potential is not limited to simple duplication or deletion of expression during development. If information that was under regulatory control of one hierarchy also is expressed within the context of other overlapping or even unrelated hierarchies, the genetic heritage of the one hierarchical order can then remain even after it is no longer coordinately expressed. Bizarre occurrences of ancestral forms are probably the best examples of this model. For instance, the enamel structure of reptilian-like teeth can still be induced in chickens (Kollar and Fisher, 1980).

Cryptic maintenance of phenotypic potential can only be one side of the coin. Changes in regulatory schemes that alter the timing of expression or even the nature of association between individual genes can result in entirely new, large-scale phenotypic experiments. The

coordinate nature of the expression of such systems will tend to mitigate (though certainly not eliminate) the negative pleiotropic effects experienced upon such changes. Goldschmidt's "hopeful monster" may not be so foreign after all.

EVOLUTION AND THE NEW GENETICS

The advent of recombinant DNA technologies has led to a profound increase in our knowledge about prokaryotic and eukaryotic gene organization—indeed, it has given birth to the "new genetics," which has expanded enormously our view of the Mendelian gene and its organizational environment. This view has been very exciting. Instead of the simplicity and order many expected, an enormously complex array of structural systems and hierarchies has been found, suggesting novel mechanisms for altering genetic information. Assuming that patterns of evolution reflect the modes of genomic variation and therefore the structural and functional organization of genetic information, the perspective provided by the "new genetics" should impose specific constraints and potentialities upon evolutionary theory. Four features of the "new genetics" have particularly important implications for our views on evolution.

1. *Genomic information is organized in a hierarchical manner whose range extends from pieces of genes to the coordination of diverse multigene families and batteries of genes.* First, genes are composed of exons that often encode the individual functional domains of multidomain proteins. Combinations of domains generate sophisticated protein molecules that exhibit synergistic cooperation of functions among the domains. Second, individual genes are frequently assembled into multigene families whose members may be virtually identical or highly diverse in character. Members of the multigene family may, in turn, be simultaneously expressed, as happens to identical multigene families such as the 5S genes, or individually expressed, as happens with informational multigene families such as the antibodies. Clearly the multigene family is a primary organizational unit for control and regulation. Finally, collections or batteries of individual genes, as well as diverse multigene families, may be functionally coordinated in a hierarchical manner to generate complex patterns of development such as that seen in the *bithorax* system. Thus, units of evolutionary information must be seen as components and orders of overlapping structural and functional organizations.

2. *The organizational features of the genome provide it with the potential for dynamic flexibility in organization.* Prokaryotic and eukaryotic genomes have mobile elements that can move around virtually at will. The movement of these elements can change dramatically the patterns of gene expression. Their dynamic nature predicts novel methods for intra- and intergenomic transmission and expansion of information. Somatically, antibody gene element rearrangements during differentiation lead to the expression of the B-cell program for development. Both of these systems contain enzymes which can alter the structure of the genome. Thus, organisms have the enzymatic machinery to change the organization of their genomes. Almost certainly, organisms will have evolved additional, as of yet unstudied, mechanisms for changing genome organization. To what extent these elaborations can operate on germline, as opposed to somatic, DNAs is a fascinating but unanswered question.

3. *These hierarchical organizations and the dynamic properties of the genome have led to a variety of mechanisms for creating rapid and extensive variation of phenotype.* These mechanisms, once again, can operate at many levels, extending from the single subcomponents of genes to their complex hierarchical organizations. First, exons can be duplicated to create new exons with the potential for diversification and the assumption of new functions. In addition, exons can be combined together in different linear combinations to create novel genes for new multifunctional proteins generating unique combinations of synergistic functions. Second, multigene families may be duplicated in part or in their entirety to create new raw material for evolution. These duplications include the cis-attendant control mechanisms as well as the structural genes—thus very complex new gene families can be created by a single genetic event. Third, the intricate program of hierarchical gene organization and/or expression such as that seen for the *bithorax* complex can be altered—thus potentially leading to strikingly new developmental features (e.g., the placement of a haltere where wings were previously located). Thus, extremely complex organizational changes, both functional and structural, can arise from few genetic events. While the evolutionary significance of this latter class of phenomena remains at present formally unproven, it may well be that large-scale systematic and phyletic variations, such as the evolutionarily rapid increase in human brain size, or even the Cambrian explosion, will be readily understandable within this context. Also, the role of speciation and its pattern within evolution will undoubtedly be considered within this context of dramatic, often coordinated phenotypic change.

4. *Selection may operate at many different levels of gene organization*. Selfish DNA sequence elements exist that experience selective constraints distinct from those of their host genomes. Selection may operate on the unique advantages conferred by a certain exon and its corresponding domain. It may view as a selective unit the integrated functions of a multi-exon gene. Selection may operate on the overall expression of an entire multigene family, or it may respond to the dramatic changes in the regulatory patterns of complex developmental programs such as that of the *Drosophila bithorax* complex. Therefore, as our knowledge of how genetic information is maintained and transmitted grows, our concept of the evolutionary unit must broaden to include relational strategies that extend from the smallest identifiable genetic element to entire organisms.

Geneticists in the last five years have developed a strikingly dynamic picture of the eukaryotic genome. Given these emerging views, new theories of evolution must clearly specify the context in which terms like selection, competition, randomness, and fixation have evolutionary meaning. Obviously, molecular genetics will not provide all of the answers for understanding evolution—rather, it must be viewed as an integral aspect of a larger synthesis that will include the rich efforts of paleontology, taxonomy, and classical genetics; for in the last analysis, knowing the organism is essential.

Cell, Vol. 40, 225–229, February 1985, Copyright © 1985 by MIT

T Cell Antigen Receptors and the Immunoglobulin Supergene Family

Review

Leroy Hood, Mitchell Kronenberg, and Tim Hunkapiller
Division of Biology
California Institute of Technology
Pasadena, California 91125

The vertebrate immune response employs two categories of cells that react specifically with antigens: B lymphocytes react with soluble antigens and T lymphocytes react with antigens that are present on cell surfaces. The receptor molecules that mediate these specific immune responses have similar heterodimeric structures: light (L) and heavy (H) polypeptides for immunoglobulins serve as the antigen receptor for B cells, and α and β chains comprise the T cell antigen receptor. Light and heavy chains, as well as α and β chains, have variable (V) regions that recognize antigen and constant (C) regions that are involved in membrane anchoring and possibly signal transmission. Different immunoglobulin heavy chain constant regions also mediate different effector functions such as complement fixation or the degranulation of mast cells. The population of T or B cells in individual animals exhibits an extremely diverse antigen-binding repertoire due to the diversity of antigen receptors. However, individual B cells express only a single type of functional receptor and the same is probably true of T cells. Antigenic challenge leads to clonal expansion of lymphocytes that have receptors specific for that particular antigen.

The Organization and Rearrangement of Immunoglobulin and T Cell Receptor Genes

Immunoglobulin and T cell antigen receptor genes have similar organizations and undergo similar DNA rearrangements. Immunoglobulin light chains are encoded by the two unlinked gene families, κ and λ. In each family, the light chain variable regions are encoded by separate V_L and J_L (joining) gene segments while heavy chain variable regions are encoded by V_H, D_H (diversity) and J_H gene segments (Figure 1). These gene segments can join together to generate complete light and heavy chain V genes. Three different classes of T cell-specific cDNA clones, α (Chien et al., Nature *312*, 31–35, 1984; Saito et al., Nature *312*, 36–40, 1984), β (Yanagi et al., Nature *308*, 145–149, 1984; Hedrick et al., Nature *308*, 149–153, 1984), and γ (Saito et al., Nature *309*, 757–762, 1984), have been recently isolated and portions of them show similarities to immunoglobulin V, J, and C regions. In each case, the genomic DNA encoding these cDNA clones is rearranged in T lymphocytes. Amino acid sequence analyses of the α (Hannum et al., Nature *312*, 65–67, 1984) and β (Acuto et al., P.N.A.S. *81*, 3851–3855, 1984) polypeptides correlate with the α and β clones, whereas the γ class does not encode any of the T cell receptor components that have been identified by numerous monoclonal antibodies (reviewed in Immunol. Reviews *81*, 1984). The function of the γ polypeptide is unknown.

The genomic DNA encoding the β and γ chains has been studied extensively. The T cell receptor V_β region, like the V_H region, is encoded by V_β, D_β and J_β gene segments that rearrange during T cell differentiation to generate a V_β gene (Chien et al., Nature *309*, 322–326, 1984; Siu et al., Cell *37*, 393–401, 1984a). Three V_γ gene segments and three J_γ-C_γ clusters have been identified (Hayday et al., Cell, this issue). The genomic organization of the α genes is unknown, but the cDNA sequences suggest that V_α, J_α, and possibly D_α gene segments rearrange to generate V_α genes.

Several observations suggest that immunoglobulin and T cell receptor genes rearrange by similar mechanisms. First, immunoglobulin light and heavy chain gene segments and T cell receptor β (Chien et al., op. cit.; Siu et al., 1984a, op. cit.) and γ (Hayday et al., op. cit.) chain gene segments have similar recognition sequences associated with DNA rearrangement. The joining boundary of each gene segment is flanked by a highly conserved heptamer, a nonconserved spacer sequence of 12 or 23 nucleotides (denoted > and >> in Figure 1) and an A/T-rich nonamer. The length of the nonconserved spacer sequence corresponds to approximately one and two turns of the DNA helix, respectively. In both immunoglobulin and T cell receptor gene segments, a one-turn recognition sequence always joins to a two-turn sequence. Second, the joining process in these gene families is imprecise; joining may occur at different nucleotides in the junctional regions of the V, D, and J gene segments. Third, in B cells, D_H is joined to J_H before the rearrangement of V_H gene segments; incomplete V_H-D_H rearrangements have not been observed (Alt et al., EMBO Journal *3*, 1209–1219, 1984). Frequent incomplete D_β-J_β joinings occur in thymocytes, suggesting that this rearrangement is also the first step in V_β gene formation, followed later by a V_β joining to the D_β-J_β rearrangement (Kronenberg et al., Nature, in press).

Diversification of T Cell Receptor Genes

T and B lymphocytes are capable of responding to a diverse spectrum of foreign molecular structures. Four strategies are employed for B cell antigen receptor diversification: multiple germline gene segments, combinatorial joining of these gene segments, somatic mutation, and combinatorial association of the polypeptide subunits. These strategies are also employed by the β gene family.

Although only a limited number of different V_β gene segments have been characterized, there are several examples of T lymphocytes with different antigen specificities that utilize the same V_β gene segment (Patten et al., Nature *312*, 40–46, 1984; Goverman et al., Cell, in press). This implies that there is only a small repertoire of V_β gene segments or that there are biases in V_β gene usage

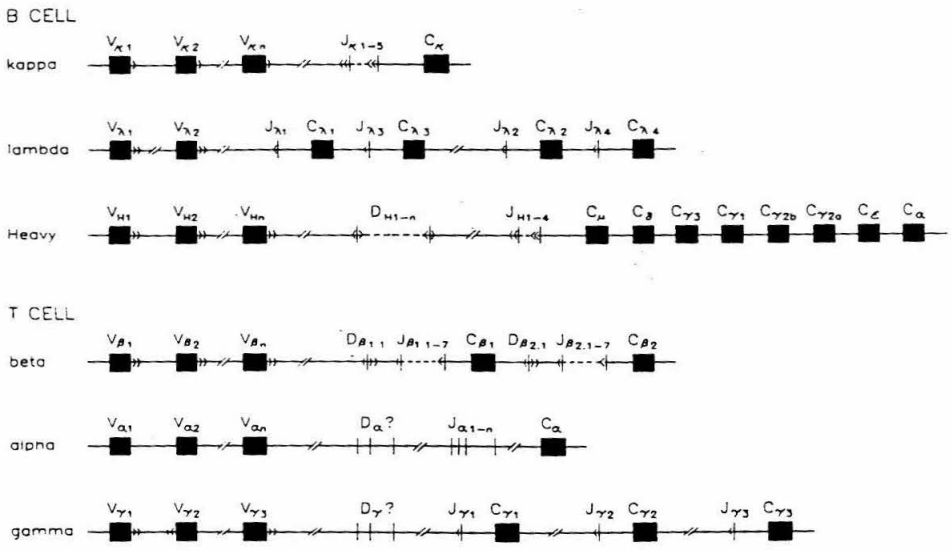


Figure 1.

such that a few gene segments from a much larger pool are used by most T cells. Because the β gene family also has twelve functional J_β gene segments, six in each cluster, and at least two D_β gene segments (Figure 1), combinatorial use of even a moderately sized V_β gene segment repertoire with the D_β and J_β gene segments would permit substantial V_β region diversity. In contrast, only one of the three V_γ and one of the three J_γ gene segments appears to be expressed in the T cells studied to date (Hayday et al., op. cit.). The mouse γ family therefore resembles the murine immunoglobulin λ gene family with regard to both J-C gene organization and the limited V gene segment repertoire (two V_λ gene segments) and, accordingly, exhibits little or no diversity due to combinatorial joining.

The β genes of T cells employ two somatic mutational mechanisms used by B cells. Junctional flexibility arises from the variability in the sites at which the gene segments may be joined (Malissen et al. Cell *37*, 1101–1110, 1984; Patten et al., op. cit.), and N-region diversification results from the random addition of nucleotides to either end of the D gene segment in the process of joining to the V and J gene segments (Kavaler et al., Nature *310*, 421–423, 1984; Siu et al., Nature *311*, 344–349, 1984b). T lymphocytes exhibit a somatic diversification mechanism not found in B cells; namely, the D_β gene segments can be joined to the V_λ gene segments in all three translational reading frames, whereas in B cells each D_H gene segment is joined in a single translational reading frame (Goverman et al., op. cit.). Somatic hypermutation, a mutational mechanism that occurs late in B cell development and can generate diversity throughout V_H and V_L genes by single base substitutions, has so far not been detected in T lymphocytes. Thus, T cell V_β diversification differs from B cell immunoglobulin diversification in that the T cell mechanisms all operate early in development at the time of V_β gene assembly, and diversify only the 3′ end of the V_β gene. The apparent lack of somatic hypermutation in β genes means that the functional β chain repertoire, and hence the T cell antigen receptor repertoire, is likely to be smaller than the immunoglobulin repertoire. The extent to which the α genes share these strategies for diversification remains to be determined.

T Cell Antigen Receptors and T Cell Function

T lymphocytes can only recognize antigen on the surface of other cells in conjunction with molecules encoded by the major histocompatibility complex (MHC). Therefore, despite the similarity between immunoglobulin and T cell receptor gene organization and rearrangement, T cell receptors differ fundamentally from immunoglobulin molecules, which can bind readily to antigens in the absence of MHC gene products. Another difference is that the T cell receptor is part of a macromolecular complex that also includes the three invariant polypeptides of the T3 antigen (Borst et al., J.B.C. *258*, 5135–5141, 1983), which may form an ion channel in response to antigen.

At least three classes of functional T cells have been defined: cytotoxic or killer (T_C) lymphocytes destroy cells displaying foreign antigen, and regulatory T cells either help (T_H) or suppress (T_S) immune responses. There is a correlation between the T cell functional class and the type of MHC-encoded molecule that is recognized in conjunction with antigen. T_C cells tend to recognize antigen in conjunction with a class I MHC molecule, while T_H cells tend to recognize antigen with a class II MHC molecule. The molecular basis for this correlation has not been elucidated, although the β genes, which are rearranged and expressed in every T_H and T_C lymphocyte examined,

are apparently not responsible. In fact, the two constant region genes, $C_{\beta 1}$ and $C_{\beta 2}$, may be expressed in either type of T cell (Kronenberg et al., op. cit.), and there do not appear to be major T_H-specific or T_C-specific biases in the use of the V_β, D_β, and J_β gene segment repertoires. Therefore, the type of β chain expressed is not correlated with either T_H or T_C functions, or with recognition of class I or class II molecules. In contrast, some but not all, apparently functional mouse T_S hybridomas delete their β gene family from both chromosomal homologues (Hedrick et al., P.N.A.S., in press; Kronenberg et al., op. cit.). It is possible that two classes of T_S cells exist: one with β genes rearranged and expressed and the other with β chains deleted. The latter class may express a different T_S-specific antigen receptor gene family.

The γ gene is transcribed preferentially in T_C lymphocytes (Hayday et al., op. cit.). This observation has led to speculation that the γ chain is part of a second receptor complex, distinct from the antigen-binding receptor, that is specific for class I MHC molecules, and is therefore intimately related to the correlation of T_C activity with MHC class I recognition.

Accessory Molecules Involved in T Cell Antigen Recognition and Function

Monoclonal antibodies against the α-β heterodimer have marked effects on T-cell function. In addition, monoclonal antibodies against four other molecules found on the surface of human T lymphocytes, including T4 (Biddison et al., J. Exp. Med. *159*, 783–797, 1984), T8 (Landegren et al., J. Exp. Med. *155*, 1579–1584, 1982), T11 and LFA-1 (Davignon et al., P.N.A.S. *78*, 4535–4539, 1981), also affect T cell function. In contrast, antibodies against other cell surface proteins have no effect on T cell function. Homologues for several of these human proteins have been found in mouse (e.g., human T4/murine L3T4, human T8/murine Lyt2) as well as in other species. The T11 and LFA-1 molecules are found on most mature T lymphocytes, while mature T cells express either T4 or T8 molecules. There appears to be a correlation between T4 expression, class II recognition and T_H function on the one hand, and T8 expression, class I recognition and T_C function on the other.

Although biochemical evidence indicates that the accessory molecules lack the structural heterogeneity expected of antigen receptors, antibodies to T4, T8, T11, and LFA-1 inhibit T cell activation by blocking an early event, such as the recognition of a cell displaying an antigenic molecule or the formation of a stable conjugate between a T_C cell and its target (Biddison et al., op. cit.; Landegren et al., op. cit.; Davignon et al., op. cit.). It has been hypothesized that the accessory molecules may be receptors that recognize nonpolymorphic structures on the target cell. The ligands that bind to these putative receptors have not been characterized, although the correlations mentioned above have led to the conjecture that the T4 molecule binds the MHC class II molecule and the T8 protein binds the MHC class I protein. The binding of ligand by the T cell antigen receptor alone may be insufficient for stable cell-cell interaction because the T cell antigen receptor has a low binding affinity or because there is a low concentration of antigen plus self MHC molecules available at the cell surface. Ligand binding by other T cell surface accessory molecules may be required to form a stable cell-cell interaction that will lead to T cell activation.

The cloning and analysis of the T8 cDNA demonstrates that this polypeptide is homologous to immunoglobulins (Littman et al., Cell, this issue; Sukhatme et al., Cell, in press) and confirms that at least some of the accessory molecules belong to what is called the immunoglobulin supergene family.

Immunoglobulin Supergene Family

A multigene family is a group of homologous genes with similar functions (Hood et al., Ann. Rev. Genetics *9*, 305–354, 1975). A supergene family is a set of multigene families and single-copy genes related by sequence (implying common ancestry), but not necessarily related in function. Sequence analysis of MHC class I (Strominger et al., Scand. J. Immunol. *11*, 573–593, 1980; Steinmetz et al., Cell *24*, 125–134, 1981) and class II (Larhammar et al., Cell *30*, 153–161, 1982; Malissen et al., Science *221*, 750–754, 1983) gene products, as well as the Thy-1 polypeptide (Williams and Gagnon, Science *216*, 696–703, 1982), suggested that these proteins belong to the immunoglobulin supergene family, named after the first genes of this family to be analyzed. In the last year, the sequences of cloned genes for the α and β chains of the T cell antigen receptor, the T cell γ chain, the poly-Ig receptor for polymeric IgA and IgM immunoglobulins (Mostov et al., Nature *308*, 37–43, 1984), and the T8 molecule have been analyzed, and all show a clear relationship with other members of the immunoglobulin supergene family. Many of the genes known to be involved in the vertebrate immune response may therefore have a common evolutionary origin.

The prototypical members of this supergene family, the immunoglobulins, are constructed primarily of polypeptide units that have a common evolutionary origin. Similarities between these homology units exist at the level of primary, secondary, and tertiary structure (Amzel and Poljak, Ann. Rev. Biochem. *48*, 961–998, 1979) as well as at the level of gene organization. Each homology unit is approximately 110 amino acids long and has a centrally-placed disulfide bridge spanning about 65 amino acids. The immunoglobulin homology units fold to generate a conserved tertiary structure, the antibody fold, which is comprised of two sheets of antiparallel beta strands. Pairs of homology units fold together to create discrete polypeptide domains (e.g., V_L-V_H, C_L-C_{H1}; see Figure 2). Thus the tertiary structure of the homology units appears to facilitate interaction with a second homology unit. Other members of the immunoglobulin supergene family are cell-surface polypeptides participating in various aspects of the vertebrate immune response and are composed of one or more immunoglobulin-like homology units (Figure 2). The degree of amino acid sequence similarity between homology units ranges from about 15% to 40% across

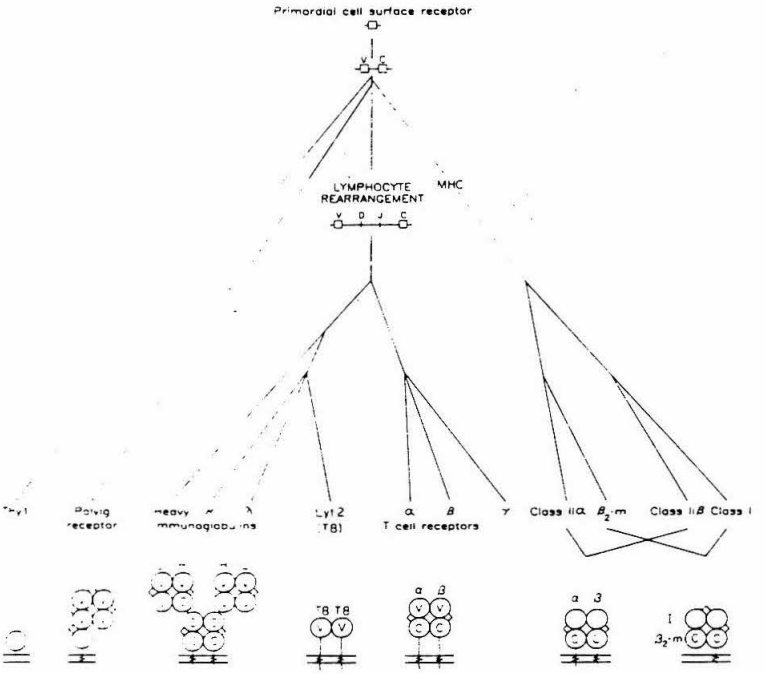


Figure 2.

gene families. At the DNA level, each homology unit is usually encoded by a separate exon, demonstrating a striking correlation between the distinct structural features of these proteins and the exon/intron structure of the corresponding gene.

Evolution of the Genes of the Immunoglobulin Supergene Family

A genealogical tree has been constructed for the single-copy genes and multigene families of the immunoglobulin supergene family by assuming that evolutionary relatedness correlates with the degree of sequence similarity among the members (Figure 2). Other features such as exon/intron structure and the ability to rearrange DNA were also assessed; since these are not easily quantified, a subjective element enters into the determination of the relative divergence times of the gene families. Nevertheless, we believe the genealogical tree depicts a likely sequence of events for the evolution of the immunoglobulin supergene family.

The primordial gene of the immunoglobulin supergene family may have encoded a cell-surface protein and would therefore have exons for a leader peptide, an immunoglobulin homology unit, and a transmembrane region. The tendency for homology units to interact suggests that this molecule may have been a homodimer or interacted with like units of other cells. A partial duplication of this gene, leading to a gene with exons encoding primordial V and C regions, was probably an early critical event in the evolution of the supergene family. Contemporary V and C homology units have little primary sequence similarity, suggesting an ancient divergence, although a few highly conserved amino acids are present in both types of units. Also, it is known that the V and C homology units of immunoglobulins have similar tertiary structures (Amzel and Poljak, op. cit.). Several features including extra beta strands in V regions and certain other conserved amino acids distinguish the V from the C homology units. This original duplication would make it possible to generate proteins with separate regions for recognition and for membrane-anchoring or effector functions.

At some point relatively early in the evolution of the immunoglobulin supergene family, the genes encoding Thy-1, the poly-Ig receptor, and the MHC class I and class II molecules may have begun to diverge from one another. The immunoglobulin homology units of the poly-Ig receptor are most like V homology units (Mostov et al., op. cit.), while one domain of the MHC polypeptides is composed of C homology units (Strominger et al., op. cit.). The Thy-1 gene may have diverged first since it contains a single immunoglobulin homology unit that is either intermediate between V and C, or somewhat more similar to a V homology unit (Williams and Gagnon, op. cit.). The Thy-1 glycoprotein is exceptional in that it is present on the cell-surface as a free homology unit and apparently does not associate either with itself or with other polypeptides. The role of Thy-1 in immune responses is unclear; it is expressed on many fibroblasts and brain cells in addition to some T cells. The poly-Ig receptor is expressed on several

glandular epithelia in liver, mammary gland, and other tissues, and it mediates the transcellular transport of polymeric immunoglobulin molecules. This receptor has five V homology units and a transmembrane region, which also has slight homology to some immunoglobulin variable regions. It is not known whether the V homology units of this receptor fold together to form domains, as is indicated in Figure 2. The class I MHC molecule is expressed on virtually all cell types. It is a heterodimer consisting of a heavy chain that is associated with β_2-microglobulin. The heavy chain has three external regions, $\alpha 1$, $\alpha 2$, and $\alpha 3$. The $\alpha 3$ region and β_2-microglobulin fold to form a domain; both are immunoglobulin C homology units. The $\alpha 1$ and $\alpha 2$ regions are either very distantly related to the immunoglobulin supergene family, or they may have an independent origin. A similar analysis applies to the class II molecules that are expressed on some lymphocytes, macrophages, and endothelial cells. The class II molecule is a heterodimer comprised of two integral membrane proteins that each have two external regions. As indicated in Figure 2, the more membrane proximal regions are homologous to immunoglobulin C homology units, while the more distal regions are of uncertain origin.

The acquisition of the ability to rearrange DNA was a critical event in the evolution of the immunoglobulin supergene family. It has been suggested that this ability may have arisen from the capture of a complex transposon by a primordial V gene (Siu et al., 1984b, op. cit.). The development of the rearrangement mechanism would present new opportunities for receptor diversification through extensive duplication of germline gene segments, through combinatorial joining of these gene segments, and through somatic mutation arising from an imprecise joining process.

The nonrearranging T8 gene is more closely related to V_κ or V_λ gene segments than to any other members of the immunoglobulin supergene family. This observation has two possible explanations. First, the T8 gene may have diverged from a V_L gene segment, lost the ability to rearrange, and acquired exons encoding a flexible hinge region, a transmembrane segment, and a cytoplasmic domain. Alternatively, the T8 gene may have diverged from a V homology unit prior to the acquisition of the rearrangement mechanism, and has, for some unknown reason, evolved in parallel with V_L gene segments.

Evolutionary Strategies in the Supergene Family

Evolution proceeds by a variety of mechanisms at different levels of gene organization in the immunoglobulin supergene family: 1) Individual nucleotides may be mutated, deleted, or inserted. 2) Exons may be deleted, duplicated, shuffled by homologous but unequal crossing over, or generated de novo through RNA splicing mutations. 3) Genes may be duplicated (or deleted) to generate multigene families that rapidly change their sizes. 4) Entire multigene families may be duplicated and adopt new functions. The duplication of a gene family creates new opportunities for evolutionary divergence and can increase the rate of evolution enormously, for not only has an array of structural genes been duplicated, but so have the cis-linked regulatory mechanisms. For example, the ability to rearrange gene segments is now found in at least six distinct gene families, presumably generated by the duplication of a primordial rearranging gene family. Thus, as the hierarchical levels of organized information become more complex, evolution can proceed at more rapid rates under appropriate selective conditions because duplication of information can occur at the successively higher levels of organization.

A variety of different cell-surface molecules that participate in the vertebrate immune response are members of the immunoglobulin supergene family; thus, in part, the evolution of this complex phenotypic trait—immunity—has proceeded within the confines of this supergene family. It now seems plausible that at least some of the other accessory molecules besides T8, and some of the other receptors for the constant region or Fc portion of immunoglobulins, will have immunoglobulin homology units. The widespread use of this unit in the immune response underscores two important features of these sequences. First, because the antibody fold may accommodate many different sequences, the homology unit can evolve rapidly to generate diverse structures (e.g., the families of different V genes of immunoglobulins and T cell receptors, the polymorphism of MHC genes). Second, homology units associate with one another to form domains; thus the duplication of gene families leads to new sets of gene families already capable of interacting. It remains to be determined whether there are genes or gene families in the immunoglobulin supergene family whose function is not associated with the immune response. Likewise, it remains to be determined whether other complex traits in eukaryotes, perhaps some of those encoded by the nervous system, will also employ the informational strategy of supergene families to evolve and elaborate their complex phenotypic traits.

Reprinted from Nature, Vol. 323, No. 6083, pp. 15-16, 4 September 1986

Immunology

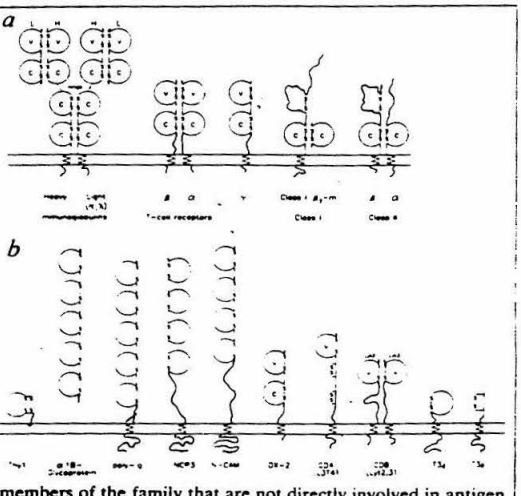
The growing immunoglobulin gene superfamily

from Tim Hunkapiller and Leroy Hood

MEMBERSHIP of the immunoglobulin gene superfamily, named for the immunoglobulin light- and heavy-chain gene families, has been growing rapidly in recent years[1]. The report by Johnson and Williams on page 74 of this issue[2] on the gene encoding a rat T-cell-associated antigen adds yet another member to the expanding list. Members of this superfamily all share a common structure called the immunoglobulin homology unit[3], a structure composed of a sequence about 100 amino-acid residues long and characterized by a centrally placed disulphide bridge that stabilizes a series of antiparallel β strands into the so-called antibody fold[4]. The variable (V) and constant (C) homology units, defined for the respective portions of the immunoglobulin chains from which they were identified, have similar but distinct three-dimensional structures. Members of the superfamily extend beyond the immune system and it appears the homology unit has played a central role in the evolution of cell–cell recognition.

The immunoglobulin gene superfamily so far includes eight multigene families and twelve single-gene representatives (see figure). The multigene families include the light (λ, κ)- and heavy-chain gene families of immunoglobulin, the α, β and γ families encoding T-cell receptors and the class I and class II genes of the major histocompatibility complex (MHC)[1]. Single-gene members include those encoding T-cell accessory molecules involved in class I (CD8)[5,6] and class II (CD4)[7,8] MHC recognition and possibly ion channel formation (T3δ, T3ε)[9,10]; a receptor responsible for transporting certain classes of immunoglobulin across mucosal membranes (poly-Ig)[11]; β_2-microglobulin, which associates with class I molecules[12]; a human plasma protein with unknown function (α1B-glycoprotein)[13]; two molecules of unknown function with a tissue distribution that includes both lymphocytes and neurones (Thy-1, OX-2)[14,15]; and two brain-specific molecules, N-CAM and neurocytoplasmic protein 3 (NP3)[16,17]. These widely divergent examples indicate the incredible evolutionary versatility of the immunoglobulin homology unit.

CD8 is a homodimer in humans and a heterodimer in rats and mice (containing the Lyt 2 and Lyt 3 chains). The primary structure of the human CD8 single chain (also denoted T8 or Leu2) and its homologue in mouse (Lyt 2) and rat (OX-8) have been previously determined and found to contain one V homology unit, most like those of light chains, as well as a connecting hinge sequence and transmembrane and cytoplasmic domains[5–7].



Domain structures of members of the immunoglobulin gene superfamily. *a*, Molecules of the immunoglobulin gene superfamily that are directly involved in antigen interaction (immunoglobulin, T-cell receptor and MHC class I and class II). The γ chain is included because of its structural similarity to the α and β chains. All these sequences are encoded by multigene families except β_2-microglobulin, included here because of its interaction with the class I heavy chain. Immunoglobulin homology-unit sequences are represented by circles held by disulphide bonds. Only disulphide bridges present in homology units are indicated. *b*, Single-copy members of the family that are not directly involved in antigen interaction. CD4 and T3δ have disulphide bridges indicated without homology-unit loops. These indicate sequences that may be descended from homology unit structures, but are now significantly divergent.

In this issue, Johnson and Williams[2] complete the molecular characterization of rat CD8 with the report of the complementary (c)DNA sequence of the Lyt 3-like chain. This chain also has one V-like homology unit, consistent with the notion that CD8 recognizes conserved determinants of class I molecules with a receptor domain similar to the paired homology unit, antigen-binding domains of immunoglobulin and the T-cell receptor[19].

The Lyt 3-like chain of rat CD8 is distinct from the Lyt 2-like chain in that is also contains a sequence highly homologous to the joining (J) segments of immunoglobulin and the T-cell receptor, particularly those of light-chains. But in contrast, the J sequence of the rat CD8 gene does not rearrange and is encoded by the same exon as the V sequence.

The V sequences of both rat CD8 chains are more similar to light-chain V regions (~30%) than they are to each other (21%). Johnson and Williams suggest that this similarity to both V and J sequences implies that the CD8 genes have descended from genes encoding a heterodimer that existed before the development of gene-segment rearrangement. However, others have suggested that the gene encoding the Lyt 2 chain of CD8 actually is an orphan V-gene segment that lost its ability to rearrange[1]. This view is supported by the extremely close linkage of human and mouse CD8 genes with immunoglobulin κ light-chain genes[20].

An intriguing alternative possiblity for the evolution of the rat Lyt 3-like gene is that it is the product of an fortuitous V–J rearrangement and translocation in a germ cell. If correct, this suggests that the sequences and perhaps even enzymes responsible for the somatic rearrangement of immunoglobulin and T-cell receptor genes can also occasionally generate rearrangments in germ cells and consequently generate new gene combinations. Less similar J-like sequences have also been proposed for the V homology units of OX-2 and human CD4 (T4)[16,8], neither of which rearrange. The suggested J sequence of CD4 can be argued against by data that indicate it is split by an intron (J. Parnes, personal communication). Interestingly, mouse CD4 (L3T4) may contain another J-like sequence elsewhere in the molecule associated with the relic of another V-like region[8]. The J-like sequences in various single-copy members of the immunoglobulin gene superfamily may represent the product of different evolu-

tionary pathways.

Though originally described as not belonging to the superfamily[10,11], re-examination of the δ- and ε-subunits of the human T-cell receptor-associated T3 molecule has led to the suggestion that both have a single *V*-like homology unit (A. Williams, personal communication). If these similarities prove convincing, it is striking that all the molecules directly associated with T-cell antigen and/or MHC recognition characterized to date (T-cell receptor, CD4, CD8 and T3) are or include members of the immunoglobulin gene superfamily.

A cDNA for α1B-glycoprotein encodes five internally repeated domains most similar to *V* homology units[14]. Its structure is very like the poly-Ig receptor, which also has five tandem, distantly related, immunoglobulin-like domains. Although membrane-bound, poly-Ig has a secretory component and α1B-glycoprotein may be the secretory component of a membrane-bound precursor[14].

The most intriguing new member of the superfamily is a developmentally regulated neuronal cell-adhesion molecule, N-CAM. Its binding is homophilic (self) and probably polyvalent[21,22]. The sequence of a partial cDNA isolated from chick embryo brain suggests that like poly-Ig and α1B-glycoprotein, N-CAM has at least four homology-unit sequences, probably arising from internal duplication. Although described as *V*-like, these units are structurally more similar to *C* homology units. As many as 500 amino-terminal residues of N-CAM remain unknown, leaving the possibility that even more immunoglobulin homology units are present. The homophilic binding function of N-CAM has been mapped to the amino-terminal portion of the molecule along with the homology units[23]. Therefore the homophilic and polyvalent nature of N-CAM may result from receptor/ligand binding analogous to the paired homology unit associations that generate the domain structures of other molecules of the superfamily. If true, this finding supports models for the origin of the immunoglobulin superfamily that suggest primordial homology units possibly mediate cell-cell interactions through homophilic associations[23].

N-CAM is a member of a new functional class within the superfamily based on a polydomain structure. Others of this group include α1B-glycoprotein and poly-Ig receptor; recent analyses suggest that even CD4 is descended from a polydomain precursor[9]. As the CD4 molecule is not known to form dimers, it may interact with class II molecules in a manner similar to the interactions of other polydomain molecules with their ligands, presumably in contrast to the heterodimeric interactions of the CD8 molecule. N-CAM is the first member of the immunoglobulin superfamily that has a brain-specific distribution.

A previously described brain protein, neurocytoplasmic protein 3, has recently been identified as a member of the superfamily (T. H., unpublished) and more recent extended analyses of this molecule suggest that it has four homology units (J.G. Sutcliffe, personal communication). Perhaps, like N-CAM, it functions in cell–cell interactions in the nervous system. Indeed, N-CAM demonstrates that members of the immunoglobulin superfamily can function outside the immune system, thus emphasizing the widespread distribution and versatility of functional characteristics of its members.

Many additional members of the superfamily are surely yet to be identified — the intriguing question is how pervasive will be the usage of immunoglobulin homology units as recognition molecules in mammalian development. □

1. Hood, L., Kronenberg, M. & Hunkapiller, T. *Cell* **40**, 225 (1985).
2. Johnson, P. & Williams, A.F. *Nature* **323**, 74 (1986).
3. Hill, R.L., Delaney, R. & Fellow, R.E. Jr *Proc. natn. Acad. Sci. U.S.A.* **56**, 1762 (1966).
4. Amzel, L.M. & Poljak, R.J. *A. Rev. Biochem* **48**, 961 (1979).
5. Littman, D.R., Thomas, Y., Maddon, P.J., Chess, L. & Axel, R. *Cell* **40**, 237 (1985).
6. Sukhatme, V.P., Sizer, K.C., Vollmer, A.C., Hunkapiller, T. & Parnes, J.R. *Cell* **40**, 591 (1985).
7. Johnson, P., Gagnon, J., Barclay, A.N. & Williams, A.F. *EMBO J.* **4**, 2539 (1985).
8. Maddon, P. *et al. Cell* **42**, 93 (1985).
9. Tourvieille, B., Gorman, S.D., Field, E.H., Hunkapiller, T. & Parnes, J.R. *Science* (in the press).
10. van der Elsen, P. *et al. Nature* **312**, 413 (1984).
11. Gold, D.P. *Nature* **321**, 431 (1986).
12. Mostov, K.E., Friedlander, M. & Blobel, G. *Nature* **308**, 37 (1984).
13. Cunningham, B.A., Wang, J.L., Berggård, I. & Peterson, P.A. *Biochemistry* **12**, 4811 (1973).
14. Ishioka, N., Takahashi, N. & Putnam, F.W. *Proc. natn. Acad. Sci. U.S.A.* **83**, 2363 (1986).
15. Williams, A.F. & Gagnon, J. *Science* **216**, 696 (1982).
16. Clark, M.J., Gagnon, J., Williams, A.F. & Barclay, A.N. *EMBO J.* **4**, 113 (1985).
17. Hemperly, J.J., Murphy, B.A., Edelman, A.M. & Cunningham, B.A. *Proc. natn. Acad. Sci. U.S.A* **83**, 3037 (1986).
18. Sutcliffe, J.G., Milner, R.J., Shinnick, T.M. & Bloom, F.E. *Cell* **33**, 671 (1983).
19. Goverman, J., Hunkapiller, T. & Hood, L. *Cell* **45**, 475 (1986).
20. Sukhatme, V.P. *et al. J. exp. Med.* **161**, 429 (1985).
21. Edelman, G. *Science* **219**, 450 (1983).
22. Cunningham, B.A., Hoffman, S., Rutishauser, U., Hemperley, J.J. & Edelman, G.M. *Proc. natn. Acad. Sci. U.S.A.* **80**, 3116 (1983).
23. Williams, A. *J. theor. Biol.* **98**, 221 (1980).

Tim Hunkapiller and Leroy Hood are in the Division of Biology 147–75, California Institute of Technology, Pasadena, California 91125, USA.

Immunological Reviews 1987. No. 100

Published by Munksgaard. Copenhagen. Denmark

L3T4 and the Immunoglobulin Gene Superfamily: New Relationships between the Immune System and the Nervous System

JANE R. PARNES* & TIM HUNKAPILLER†

INTRODUCTION

L3T4 (mouse CD4) is a cell surface glycoprotein expressed on most thymocytes and on the predominant subset of peripheral T lymphocytes. Mature T lymphocytes can be divided into two mutually exclusive populations based on their expression of either L3T4 or the alternative T cell surface marker, Ly-2,3. Traditionally, these molecules have been considered to characterize T cells based on their function, with L3T4 expressed primarily on helper and/or inducer T cells and Ly-2,3 primarily on cytotoxic and/or suppressor T cells. While this classification is generally true, there are many exceptions, and the expression of these distinct surface proteins has been found to correlate best with the class of major histocompatibility complex (MHC) molecule recognized by the T cell (Dialynas et al. 1983, Swain 1983). L3T4 cells recognize or are restricted by class II MHC molecules (H-2 K, D or L), while Ly-2,3 T cells recognize or are restricted by class I MHC molecules (IA or IE). Monoclonal antibodies (MAbs) specific for L3T4 or Ly-2,3 inhibit the function of T cells which bear these proteins and have T cell receptors which recognize the appropriate MHC molecule (Hollander et al. 1980, Sarmiento et al. 1980, MacDonald et al. 1981, Dialynas et al. 1983). The best functional data to date suggest that at least one function of these T cell surface proteins is to increase the avidity of the interaction between the T cell and antigen-presenting or target cells (MacDonald et al. 1983, Dialynas et al. 1983, Swain 1983, Marrack et al. 1983, Greenstein et al. 1984, Dembic et al. 1987, Gabert et al. 1987). It has been hypothesized that this is accomplished by binding

* Division of Immunology, S-021, Department of Medicine, Stanford University Medical Center, Stanford, CA 94305, and † Division of Biology, California Institute of Technology, Pasadena, CA 91125, U.S.A.

of L3T4 or Ly-2,3 to nonpolymorphic regions on class II or class I MHC molecules, respectively (MacDonald et al. 1983, Dialynas et al. 1983, Swain 1983, Marrack et al. 1983, Greenstein et al. 1984, Dembic et al. 1987, Gabert et al. 1987). Although there is no direct evidence for such a receptor-ligand interaction, much of the available functional data would be explained by such a role. It has also been suggested that these molecules might function, at least under certain circumstances, by transmitting a negative signal to the T cell (Hunig 1984, Wassmer et al. 1985, Tite et al. 1986). This hypothesis has been used to explain examples in which MAbs specific for these surface markers inhibit lectin-stimulated T cell activation or function in the absence of the appropriate class of MHC molecule. However, there is currently no convincing data showing that such transmission of a negative signal is a normal physiological role for L3T4 and/or Ly-2,3.

L3T4 and Ly-2,3 have been called T cell differentiation antigens because of their differential expression during the development of thymocytes into mature T cells. A small population of thymocytes (< 5% of the total) expresses neither of these proteins and such cells are called "double negatives" (Ceredig et al. 1983, Mathieson & Fowlkes 1984). These cells are believed to include the earliest population of T cell precursors in the thymus. The major population of thymocytes, representing about 80% of the total, expresses both L3T4 and Ly-2,3. These cells are referred to as "double positives" and are found predominantly in the thymic cortex (Dialynas et al. 1983, Scollay et al. 1984). The most mature thymocytes, the so-called "single positives", are found mostly in the thymic medulla and express either L3T4 (about 11%) or Ly-2,3 (about 4%), similar to mature T cells in the periphery (Scollay et al. 1984). It is still unknown whether the single positive cells arise from the double positives or as a separate lineage directly from double negative cells. What is known is that most of the double positive cells die in the thymus, and at least the vast majority of T cells which emerge from the thymus are single positives (Scollay et al. 1984).

STRUCTURE OF L3T4

The L3T4 molecule was first described by Dialynas et al. (1983), who generated a rat MAb, GK1.5, which binds to a mouse T cell surface protein. Studies with MAb GK1.5 indicated that the mouse molecule is similar in distribution and structure to the human CD4 molecule (Dialynas et al. 1983). L3T4 was shown to be a single-chain glycoprotein of apparent molecular weight 52 000 daltons (Dialynas et al. 1983). We recently isolated cDNA clones encoding L3T4 using a human CD4 probe (Tourvieille et al. 1986). Transfection of these clones in a cDNA expression vector into L cells or T cells yields a cell surface molecule which binds to 2 MAbs specific for distinct epitopes of L3T4 (unpublished results). We deduced the amino acid sequence of the L3T4 protein from the

nucleotide sequence of the cDNA clones (Tourvieille et al. 1986) (Fig. 1). The amino-terminus of the mature protein has been determined by Classon et al. (1986) by amino acid sequencing. The mature L3T4 polypeptide chain consists of 431 amino acids with a predicted molecular weight of 47 918 daltons (Tourvieille et al. 1986). The nascent polypeptide chain is preceded by a 26 amino acid signal peptide which is presumably cleaved during passage through the membrane of the endoplasmic reticulum. L3T4 is predicted to have 368 amino acids external to the cell, a 25 amino acid hydrophobic transmembrane segment and a 38 amino acid extremely basic cytoplasmic tail. There are four potential sites of N-linked glycosylation, at residues 161, 272, 297 and 366. The mature protein has a total of 13 cysteine residues. Six are external to the cell, four are in the membrane and three are in the cytoplasmic tail. Analysis of the disulfide bridges in the mouse and sheep proteins has suggested that all six of the cysteines in the extracellular portion of the molecule are involved in intrachain disulfide bonds to adjacent cysteine residues (Classon et al. 1986). For mouse L3T4 this would imply disulfide loops between cysteines 16 and 86, 133 and 162, and 302 and 344.

When compared to sequences in the Protein Information Resource (PIR) data library, the amino-terminal portion of L3T4 (approximately 100 amino acids) was found to be related by sequence to immunoglobulin (Ig) variable (V) regions, particularly to those of kappa light chains (up to 35% identical) (Tourvieille et al. 1986). This domain contains the two cysteine residues characteristic of members of the Ig gene superfamily (discussed below). Various computer predictions of secondary structure potential were also very similar for this amino-terminal domain of L3T4 and for V regions of Ig light and heavy chains, as well as those of the T cell receptor (Tcr) (data not shown). These findings suggest that the V-like domain of L3T4 may fold in very similar fashion to Ig V regions. There is a second V-related domain in L3T4 immediately downstream of the first (Tourvieille et al. 1986) (Fig. 1). However, this domain (V′) is severely truncated, maintaining ony a sequence related to the carboxy-terminal half of Ig V regions. Comparisons to other V's indicate that this foreshortened domain is most similar to the analogous region of the amino-terminal or primary V-like domain of L3T4, suggesting that there may have been a duplication of at least part of the primary V-like domain (see below). Although the L3T4 gene does not rearrange, each of these V-like domains of L3T4 is followed by a sequence which bears similarity to Ig joining (J) segments (Tourvieille et al. 1986) (Fig. 1). The J-like sequence following V′ (denoted J′) is the more convincing of the two based on sequence comparisons, and it is followed by an intron at the appropriate place for Ig and Tcr J sements (Fig. 2). In contrast, the sequence of the J-like segment following the amino-terminal V-like domain (denoted J) is less closely related to Ig J segments, less conserved between species, and is interrupted at the DNA level by an intron (Fig. 2). However, the evolutionary relationship of either region to Ig J's is still uncertain.

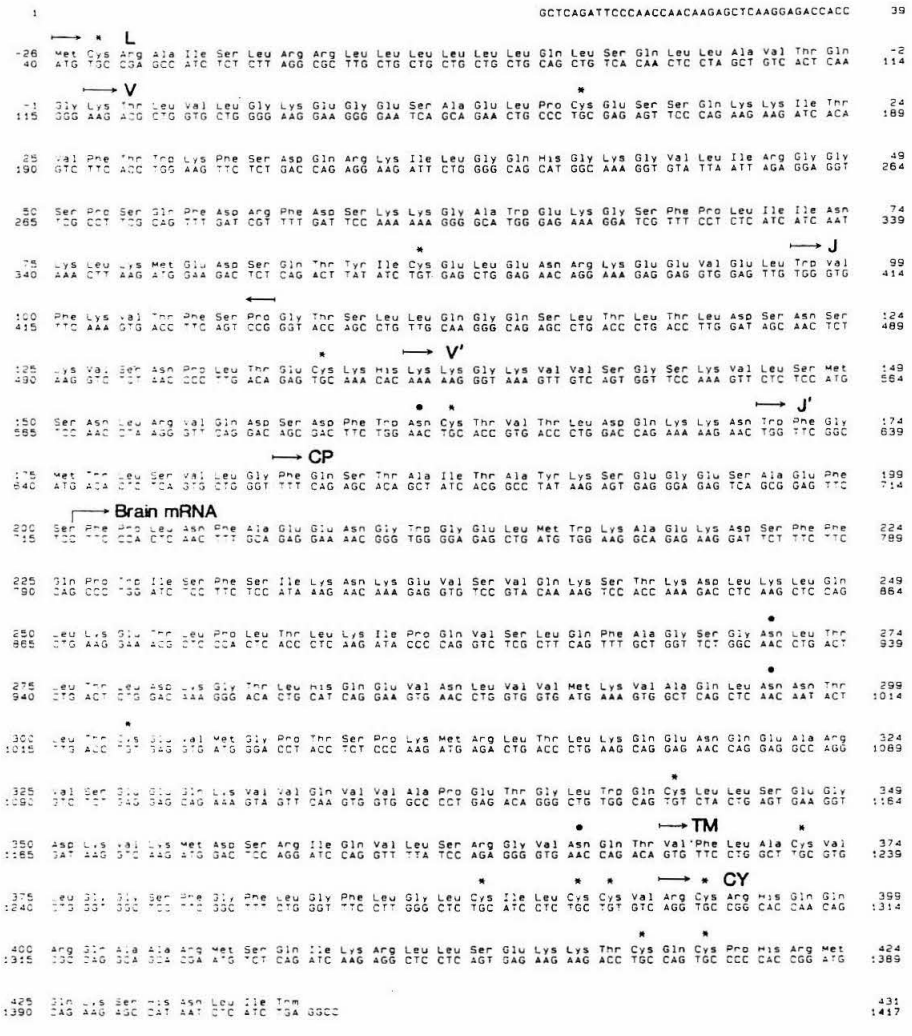


Figure 1. Nucleotide and deduced amino acid sequence of L3T4 cDNA. The nucleotide sequence of the L3T4 cDNA is shown with the amino acid translation above. The number of the first and last amino acid and nucleotide in each line is indicated in the left and right margins, respectively. Predicted protein domains are labeled: L, leader; V and V′ sequences homologous to Ig V regions; J and J′, sequences analogous to Ig J segments; CP, connecting peptide; TM, transmembrane region; and CY, cytoplasmic tail. Cysteine residues are marked by asterisks and potential N-linked glycosylation sites by closed circles. The approximate (±10 bp) start of the L3T4 transcript in brain is indicated by a horizontal arrow.

L3T4 AND THE Ig GENE SUPERFAMILY

We have referred to the carboxy-terminal half of the external domain of L3T4 as a connecting peptide because its function is unknown and it is not closely related to other known proteins. It has been suggested (Clark et al. 1987) that this portion of the molecule may be constructed of two additional, albeit degenerate V-like domains. The more membrane-distal of these two has no cysteine residues, while the membrane-proximal domain has a foreshortened disulfide loop (spanning only 43 amino acids in the mouse protein). If this connecting peptide region is derived from ancestral precursors of the Ig gene superfamily, it is clear that it has diverged far more than the amino-terminal portion of L3T4. Others have suggested that this region is more closely related to other, non-Ig-related proteins (B. Jameson, personal communication). However, statistical arguments are not completely convincing for either of these postulated relationships. More diverse phylogenetic comparisons of L3T4 homologs may help determine the true history of this region.

STRUCTURE OF THE L3T4 GENE

L3T4 is encoded by a single gene on mouse chromosome 6 (Tourvieille et al. 1986, Field et al. 1987). Using cDNA clones as probes, we isolated a series of overlapping genomic clones containing the entire L3T4 gene. Analysis of these clones indicated that the L3T4 gene spans 26 kb of DNA and is composed of 10 exons (Gorman et al. 1987) (Fig. 2). We determined the nucleotide sequence of the exons, the intron/exon borders and at least a portion of each intron (Gorman et al. 1987). One unusual, although not unique feature of this gene is the presence of an intron (8.6 kb) within the 5′ untranslated region. However, the most striking

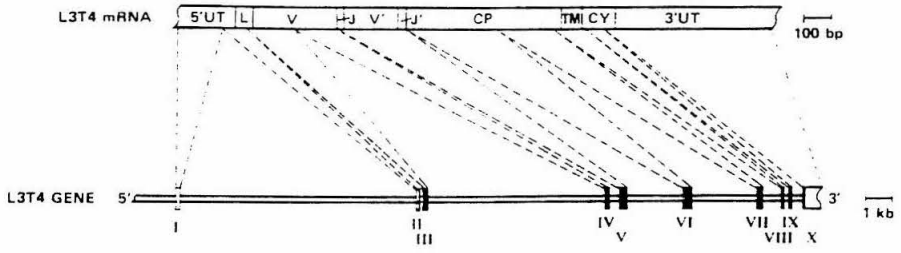


Figure 2. Structure of the L3T4 gene. The structural organization of the L3T4 gene is shown in the bottom half of the figure. Exons are indicated by boxes and numbered with Roman numerals. Protein coding regions of the exons are indicated by shaded boxes, and untranslated regions by open boxes. Introns and flanking sequences are denoted by open bars. Dotted lines indicate where the exons are represented along the structure of the L3T4 mRNA, shown in the top half of the figure. Protein coding sequences are labeled as in Fig. 1. 5′UT, 5′ untranslated region; 3′UT, 3′ untranslated region.

and unexpected finding was that a large intron (6.4 kb) divides the sequence encoding the amino-terminal V-like domain into two exons (exons III and IV) approximately half-way through the predicted protein domain (Gorman et al. 1987, Littman & Gettner 1987). With this exception, the remainder of the intron/exon structure correlates reasonably well with the predicted protein domains (Fig. 2), a feature characteristic of members of the Ig gene superfamily. Interestingly, the connecting peptide region, which may contain two additional domains ancestrally related to Ig V regions, is appropriately split between two exons. As is true for other members of the Ig gene superfamily, introns interrupt codons between the first and second nucleotides (1/2 codon split) in all cases except between the exons encoding the evolutionarily unrelated cytoplasmic tail.

L3T4 mRNA IS EXPRESSED IN BRAIN

L3T4 protein has only been identified in thymocytes and mature T lymphocytes. Although the human and rat homologs are also expressed in macrophages and the related Langerhans cells (Wood et al. 1983, Moscicki et al. 1983, Jefferies et al. 1985), this does not appear to be true in the mouse (Crocker et al. 1987). The availability of L3T4 cDNA clones led us to examine the cellular distribution of L3T4 mRNA by Northern blot analysis (Tourvieille et al. 1986). We identified a 3.7 kb L3T4 mRNA species in mouse thymus, spleen, lymph node and in T cell lines which express L3T4 protein (Fig. 3). We did not see expression of this mRNA in T cell lines which do not express surface L3T4 or in mouse liver, B cell lines, a macrophage cell line, F9 teratocarcinoma cells or a rat glial cell line. Although we detected a very small amount of this mRNA species in kidney, it is unclear whether that represents true expression by kidney cells or rather the presence of contaminating T cells. In contrast, we found that mouse brain expresses low levels of a smaller mRNA species (2.7 kb) that hybridizes to the L3T4 probe (Fig. 3). We did not see a similar mRNA species in any other tissue or cell line examined. Even lower levels of the typical 3.7 kb L3T4 mRNA are found in mouse brain, but this result is difficult to reproduce. Maddon et al. (1986) examined different sections of mouse brain for expression of mRNA which hybridizes to an L3T4 cDNA probe. They identified the shorter, brain-specific transcript in forebrain but not hindbrain. They also cited unpublished data showing that it is present in the cortex and most abundant in the striatum, but absent from the cerebellum, brain stem and spinal cord. We have begun to examine which cells in brain express this L3T4 transcript. By *in situ* hybridization we have found large amounts of mRNA which specifically hybridizes to an L3T4 probe in primary cultures of mouse brain astrocytes (manuscript in preparation). We do not yet know whether L3T4 mRNA is also present in neurons or in other glial cells.

We had shown that L3T4 is encoded by a single gene in the mouse genome

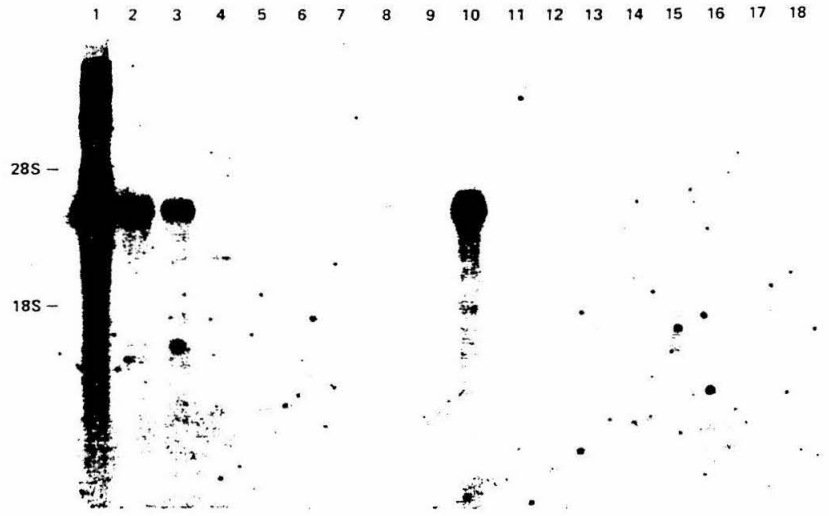


Figure 3. Northern blot analysis of L3T4 mRNA expression. A Northern blot of total RNA (10 μg) from the indicated sources was hybridized to a ^{32}P-labeled L3T4 cDNA probe and exposed to x-ray film. All RNA was of mouse origin unless otherwise indicated. RNA sources were: 1, thymus; 2, lymph node; 3, spleen; 4, brain; 5, kidney; 6, liver; 7, rat glial cell line C6; 8, thymoma cell line KKT2; 9, no sample; 10, thymoma cell line VL3/1; 11, thymoma cell line 1112; 12, thymoma cell line MBL2; 13, thymoma cell line R1.1R/TLIII 7X.6; 14, B-cell line Bal 1; 15, B-cell line 225; 16, F9 teratocarcinoma cell line; 17, macrophage cell line WEHI-3; 18, WEHI-3 after 48 h of induction with cell-free supernatant from concanavalin A-treated mouse spleen cells. The migration positions of 28S and 18S ribosomal RNAs are indicated in the left margin.

(Tourvieille et al. 1986), so it was clear that the smaller transcript seen in brain must be a product of the same gene. We demonstrated that this transcript was polyadenylated, and therefore unlikely to be a degradation product of the normal L3T4 transcript (Tourvieille et al. 1986) (Fig. 4). We therefore examined the structure of this brain-specific transcript both by hybridization of Northern blots to cDNA probes defining different regions of the L3T4 gene, and by S1 nuclease analysis (Gorman et al. 1987) (Fig. 5). We found that the 2.7 kb transcript only contains sequences from the 3′ untranslated region and the 3′ half of the L3T4 coding sequence. By Northern blot analysis we could not detect any hybridization of this mRNA species to probes containing the 5′ untranslated region (data not shown). We mapped the apparent start site of this transcript to the sequence encoding amino acid 200 (±10 nucleotides) (Gorman et al. 1987) (Figs. 1 and 6). This site does not correlate with an exon/intron junction, but rather lies within

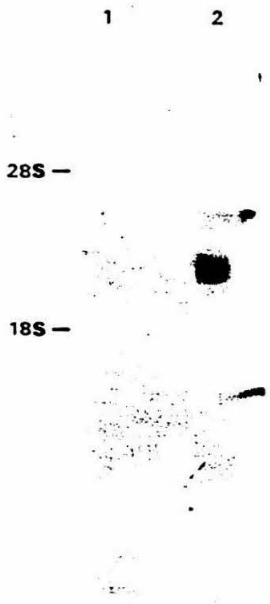


Figure 4. The transcript of L3T4 in mouse brain is polyadenylated. Poly(A)$^-$ (lane 1) and poly(A)$^+$ (lane 2) fractions of total mouse brain RNA were isolated and a Northern blot prepared. The blot was hybridized to a ^{32}P-labeled L3T4 cDNA probe and exposed to x-ray film. The migration positions of 28S and 18S ribosomal RNAs are indicated in the left margin.

protein coding sequence of exon VI (Figs. 1 and 6). The brain-specific mRNA species then continues colinearly with the T cell form of L3T4 transcript. Although we cannot completely exclude the possibility of a small amount on non-contiguous sequence at the 5′ end, these findings indicate that the brain-specific transcript cannot result from a simple alternative pattern of mRNA splicing, and rather suggest that it is most likely the product of an alternative transcriptional start site. In accord with this conclusion we found the sequence TATAA located 32 bp upstream of the predicted start site of the 2.7 kb transcript (Fig. 6). This sequence, which is also located within exon VI, may represent a promoter that is regulated specifically in brain. The first ATG (methionine codon) downstream from the predicted transcriptional start site is located 43 bp more 3′ (Fig. 6), and would be the predicted initiation site for translation. This ATG is in the same frame as that of the full-length T cell form of L3T4 mRNA. The predicted protein from the brain transcript would therefore be a truncated form of L3T4 beginning

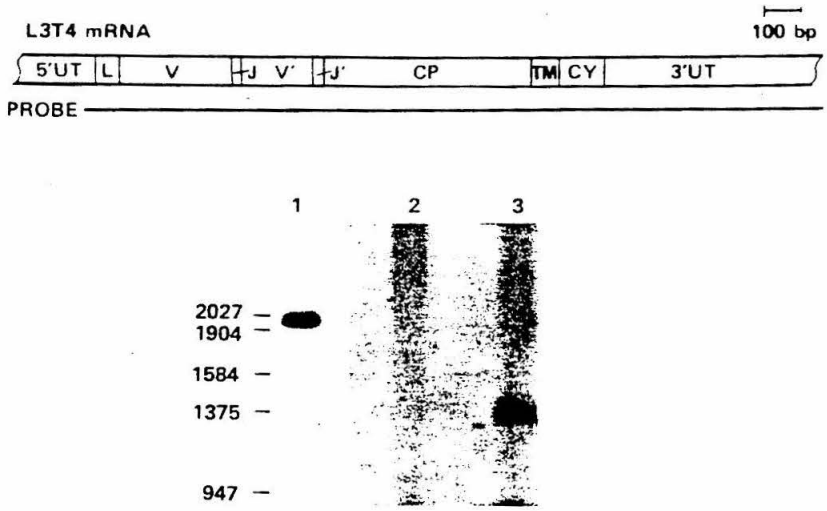


Figure 5. S1 nuclease analysis of L3T4 transcript in mouse brain. An S1 analysis of the L3T4 mRNA in mouse brain was performed as described (Gorman et al. 1987). The single-stranded probe used for protection consisted of the anti-sense strand of an L3T4 cDNA clone of approximately 2 kb in length, beginning within the leader sequence. (Note: only that part of the 3′ untranslated region contained in our cloned cDNAs is illustrated in the mRNA diagram). This probe protected a fragment equal to its full-length in thymus RNA (lane 1). The probe did not protect a fragment in liver RNA (lane 2), but did protect a fragment of about 1.4 kb in brain RNA (lane 3). The latter protected fragment could be shown to extend linearly from the 3′ end of the probe (Gorman et al. 1987).

at amino acid 214 (within what we have called connecting peptide sequence) and continuing through the normal cytoplasmic tail. It would be missing the normal signal peptide and all of the sequences that are most similar to Ig V regions and J segments, i.e., the portion of the mature protein we believe most likely to be involved in receptor function. Without a typical signal sequence, it is not known what the fate of such a protein would be. It is possible that another internal sequence could act in this capacity, or that the protein does not get expressed on the cell surface. In the latter instance it might either be maintained within the cell or degraded. In any event, the predicted protein would certainly not function in the same manner as the full-length L3T4 molecule on T cells. We have examined fixed mouse brain sections with MAbs specific for two distinct epitopes of L3T4 and have failed to see any specific staining (unpublished results). Only one of these MAbs could have any possibility of binding to the product of the 2.7 kb transcript since the other, GK1.5, is specific for a determinant within the amino-

Figure 6. Location of the transcriptional start site of the L3T4 mRNA in mouse brain. The nucleotide sequence of exon VI and parts of the preceding and following introns is shown with the translated amino acid sequence of the exon indicated above. The numbers of the first and last amino acid in each line are indicated in the left and right margins, respectively. The start site of the L3T4 mRNA in brain is indicated by a horizontal arrow. The TATAA sequence which may act as a promoter for this transcript is overlined. The ATG (methionine codon) that may be used as a start site for translation of this brain transcript is boxed. The GT (donor) and AG (acceptor) sequences at the splicing junctions are underlined.

terminal 172 amino acids (B. Tourvieille and J. R. Parnes, unpublished results). Although the second MAb recognizes a non-overlapping more membrane-proximal determinant (B. Tourvieille and J. R. Parnes, unpublished results), it is possible that the folding of the truncated protein (if stably synthesized) eliminates this epitope even if the linear sequence is entirely present. We also found that the primary mouse astrocytes which contained L3T4-specific mRNA by *in situ* hybridization did not stain with either of these MAbs (manuscript in preparation). New MAbs specific for the carboxy-terminal portion of L3T4 may be required to determine whether cells in brain contain a stable protein product of the 2.7 kb transcript of the L3T4 gene and, if so, what the function of that protein product might be.

The finding of L3T4 mRNA in brain became even more intriguing following evidence from a variety of laboratories indicating that the human homolog, CD4, is the receptor for human immunodeficiency virus (HIV), the retrovirus that causes acquired immune deficiency syndrome (AIDS) (Dalgleish et al. 1984, Klatzman et al. 1984, McDougal et al. 1986, Maddon et al. 1986). Central nervous system involvement, especially a subacute encephalopathy, occurs with high frequency in patients with AIDS (Snider et al. 1983, reviewed in Johnson & McArthur 1986). Maddon et al. (1986) indeed found CD4 transcripts to be present in human cerebral cortex mRNA. In contrast to findings in the mouse, human brain contains not only a shorter transcript, but also easily identifiable quantities

of "full-length" CD4 mRNA that comigrates with human T cell CD4 transcripts (Maddon et al. 1986). Although the equivalent TATAA sequence is conserved in the human gene (Maddon et al. 1985), it is not yet known whether the shorter CD4 transcript in human brain is analogous to that seen in the mouse. If it is, it could not be translated into a protein similar to the predicted translation product in the mouse, because the equivalent methionine codon is absent. (This methionine codon is also absent in the rat CD4 homolog [Clark et al. 1987]). It is most likely that receptors for HIV in human brain result from translation of the full-length CD4 transcript. MAbs specific for CD4 have been found to stain human brain sections (Pert et al. 1986, Funke et al. 1987). Since human macrophages are known to express CD4, they are most probably the predominant class of $CD4^+$ cells in brain. This is supported by data demonstrating the presence of HIV mainly within mononuclear and multinucleated macrophage-like cells in the brains of patients with AIDS involving the central nervous system (Epstein et al. 1985, Koenig et al. 1986, Wiley et al. 1986). However, one recent study suggested that CD4 may also be present on neuronal and glial cells in human brain (Funke et al. 1987), and on rare occasions HIV has been observed in astrocytes in the brains of AIDS encephalopathy patients (Epstein et al. 1985, Wiley et al. 1986). The physiological role (if any) of CD4 on these non-T cells is not known.

EVOLUTIONARY CONSIDERATIONS

Ig gene superfamily

Over the last several years the number of known members of the Ig gene superfamily has grown dramatically beyond the actual immune receptors, Ig and Tcr (for review, see Hunkapiller & Hood 1988). L3T4 is representative of a class of immune-related members that includes class I and class II MHC antigens, β_2-microglobulin, Ly-2,3 (CD8), T3 (CD3), poly-Ig receptor (PIR) and the gamma Fc receptor (FcR). The most recently defined class of members includes a series of molecules with growth factor properties, including the receptor for platelet-derived growth factor (PDGF-R), the receptor for macrophage colony stimulating factor (CSF-1R) and the oncogene v-kit. However, potentially the most interesting class of new member molecules consists of those which are expressed on cells of the nervous system. Among these are a series of cell adhesion molecules: neuronal cell-adhesion molecule (N-CAM), myelin-associated glycoprotein (MAG) and P_0, a peripheral myelin-associated molecule. Yet what makes this class particularly intriguing are those representatives, like L3T4/CD4, that are expressed by both lymphocytes and nervous system tissues, including two molecules of unknown function, Thy-1 and OX-2, as well as, possibly, MAG (see below).

A molecule is considered homologous to the Ig gene superfamily if it maintains at least one Ig "homology unit" (Hill et al. 1966). As described, a homology unit

is a sequence of about 100 residues in length which has a series of conserved residues significant for maintaining a characteristic tertiary structure, referred to as the "antibody fold" (see Amzel & Poljak 1979). This structure is constructed from a series of anti-parallel beta strands that assume a sandwich-like configuration. The two planes are primarily stabilized by an invariant disulfide bond between an internal pair of cysteines. Homology units are often described as V-like or C-like depending on their similarity to Ig V or Ig constant (C) region sequences. Each type has a specific, although overlapping set of conserved residues. However, C homology units lack an internal pair of beta strands relative to V units. Structurally, these strands are found between the layers of the sandwich, and their absence or presence decreases or increases, respectively, the distance between the two cysteine residues. Various comparative analyses indicate that L3T4 has a definite V-like homology unit (data not shown).

Although comparison to other members of Ig gene superfamily indicates that the L3T4 homology unit shares the greatest similarity with the V region sequences of Ig and Tcr (Tourvieille et al. 1986), this region is not as similar to true V regions as the homology units of Ly-2 and Ly-3 (Zamoyska et al. 1985, Nakauchi et al. 1985, Gorman et al. 1987). These latter can most easily be viewed as "orphan" light chain genes that have lost their ability to rearrange further (Hunkapiller & Hood 1986). The presence of a single light-chain-like V region in both, a well conserved J sequence in Ly-3 and the close linkage to the kappa locus of both genes in mouse (Gottlieb 1974) and at least the CD8 gene in humans (Sukhatme et al. 1985) all tend to support this notion.

Although it may have an analogous function to Ly-2,3, L3T4 has an organization more akin to the polydomain structures of the Ig-binding proteins (PIR and FcR), the neural cell adhesion molecules (N-CAM amd MAG), and the growth factor members (PDGF-R and CSF-1R). Each of these are surface-bound molecules with four or five extracytoplasmic Ig-like domains. However, this common structural motif does not necessarily imply direct evolutionary relationships. Only N-CAM and MAG have any hint of a direct correspondence between individual domains. Rather, it seems that tandem duplication of homology units is a frequently advantageous occurrence. The general correlation between exons and protein domains and the conserved 1/2 codon split between exons in members of this family should facilitate the production of a functional product from such a genetic event. In fact, a polydomain organization is the most common motif seen within the Ig superfamily. This prevalence, no doubt, reflects the inherent functional properties of homology units, primarily their tendency to specifically interact or bind with other homology units (see Hunkapiller & Hood 1988). Polydomain structures may be important to increase valency or to produce internal pairs of homology units necessary to define receptor structures. It is unclear, then, if the differences in organizational structure between Ly-2,3 and L3T4 imply significant functional distinctions or are merely the manifestation of

different evolutionary pathways to convergent functional properties. Epitope mapping of CD4 by MAbs indicates that the V and V′ domains may interact to form a "dimeric" structure analogous to the Ly-2,3 dimer (B. Jameson, personal communication).

Evolution of L3T4

Although L3T4 may or may not be the descendant of a four-domain sequence, it certainly has derived from a multiple-domain sequence. As is true for most of the other polydomain sequences in the superfamily, L3T4 shows no domain-to-domain correspondence to any of the other polydomain sequences. However, there are interesting similarities between the sequence of L3T4 and those of other, non-immune members of the superfamily. Of particular note are the similarities between L3T4 and N-CAM that may indicate a more direct evolutionary relationship between these two molecules (Matsunaga & Mori 1987). Except for MAG, to which it is most closely related, N-CAM has its greatest degree of similarity between its first homology unit and that of L3T4, although this similarity is still less than that between L3T4 and Ig light chains. N-CAM, like L3T4, also has a rare intron in the 5′ untranslated sequence (Owens et al. 1987). Most significantly, however, N-CAM is the only other member of the Ig gene superfamily with an intron splitting its Ig-like domains, and this is true for all five of its domains (Owens et al. 1987). The occurrence of introns in approximately the same position in each of the N-CAM homology units is again supportive of the scenario of tandem expansion of the number of units through homologous, unequal crossing over. (A more obvious example of this phenomenon is found in the most recently described member of the superfamily, carcinoembryonic antigen [Thompson et al. 1987, Zimmerman et al. 1987]). It has been argued that the presence of these introns implies that L3T4 and N-CAM are representative of the most primordial Ig superfamily members and that such sequences support the notion that homology units themselves are derived from the tandem duplication of a primordial "half-domain" (Littman & Gettner 1987).

As intriguing as these arguments are, there are important caveats. Interestingly, although both N-CAM and L3T4 are more similar to V homology units when measured by number of identical residues, N-CAM domains have a definite C-like nature with a particularly short distance between the two cysteines when their β structures are predicted. In contrast, the primary homology unit of L3T4 has no particular C-like nature. N-CAM homology units may, in fact, represent more generic or primordial homology units which are neither V- nor C-like. Many of the other polydomain sequences, including MAG and the growth members, also have generic homology units. It's most likely that they represent, as a group, the most primitive motif (Hunkapiller & Hood 1988). The strong V-like character of L3T4 definitely argues against its inclusion in this group. Also,

the presence in L3T4 of homology units with (i.e., the first domain) and without (i.e., the second domain, and if one accepts them as homology units, the third and fourth domains) a dividing intron may more easily argue that the intron in the first domain has been inserted after the divergence of L3T4 from other members of the superfamily. A similar event could have happened independently during the history of N-CAM. It will be most interesting to learn the exon/intron structure of MAG, which is less distantly related to N-CAM.

As mentioned above, the V unit remnant that constitutes the second domain of L3T4, V′, is more similar to the primary domain of L3T4 than it is to any other Ig superfamily member (Tourvieille et al. 1986). This implies a tandem duplication of either part or all (with subsequent loss of much of the sequence) of the primary unit. The exon/intron structure of the first domain may have played an important role in this event. The sequence around the first cysteine of V′ is no more similar than random to the equivalent region of other Ig superfamily members in general, or the primary (first) domain of L3T4 in particular. It may be that only the second exon of the first domain was duplicated, or that a subsequent deletion removed all of the first exon of the second domain. The 1/2 codon split between exons would keep such a transcript in frame, facilitating the development of the current structure. This implies that the first cysteine of this domain is not a direct descendant of that of the primary domain.

Immunoglobulin gene superfamily and the nervous system

It is still unclear whether L3T4 and/or human CD4 are expressed as functional proteins on brain-specific cells. The possible role in brain of such proteins, either truncated or full-length, is open to conjecture. However, the possibility that they might be functional highlights the larger issue of why the same or related molecules are present on both immune and nervous system cells. Evolutionarily, the finding that many immune system proteins and certain nervous system proteins are distantly related members of the same Ig gene superfamily is most suggestive of a cell-adhesion function of the primordial members of this family (Williams 1982). However, this does not explain those instances in which the same molecule is expressed on both types of cells. Besides possibly L3T4 or CD4, Thy-1 and OX-2 are also present on cells of both the immune system and the nervous system, including thymocytes and neurons (as well as a variety of other cells for Ox-2 and Thy-1). Thy-1 is organized essentially as a single V-like homology unit attached to the cell surface by a glycophospholipid moiety (Cohen et al. 1981, Tse et al. 1985). OX-2, on the other hand, looks very much like a Tcr chain with one V- and one C-like region, including a J-like connector (Clark et al. 1985). The functions of both of these proteins are unknown, and neither sequence is particularly similar to any other superfamily member.

Without a defined function, these members do not much enlighten our perspec-

tive on the significance of coexpression between these two cell systems. However, a recent discovery concerning MAG may eventually prove interesting in this regard. MAG appears to be involved in neuron-glial and glial-glial cell interactions leading to myelination and the maintenance of periaxonal space, possibly in a manner similar to the homophilic adhesion interactions of N-CAM (see Arquint et al. 1987). Recently, a protein has been described on cells of both the T and B cell lineages with the serological and biochemical characteristics of MAG (Peault et al. 1987). This protein is conserved phylogenetically and is regulated in a tissue- and developmentally-specific manner. If this protein is MAG, it is suggested that its cell adhesion properties may be involved in the "homing" process of cells of both the nervous and immune systems (Peault et al. 1987). The tendency of Ig superfamily members to associate through their homology units raises the interesting question of whether the cell-cell interactions involved in these homing processes might be facilitated or regulated by interactions between MAG and some of the other superfamily members that are also expressed on both cell systems, e.g. Thy-1, OX-2, and possibly L3T4 or CD4.

SUMMARY

L3T4 is a mouse cell surface protein expressed on most thymocytes and on the subset of mature T cells that recognizes class II MHC molecules. Its primary function on T cells is most likely that of increasing the avidity of the interaction between T cells and antigen-presenting or target cells. It may accomplish this by binding to a nonpolymorphic region on class II MHC molecules. The cDNA and gene encoding L3T4 have been isolated and sequenced. Analysis of the amino acid sequence predicted by the nucleotide sequence indicates that L3T4 is a member of the Ig gene superfamily. It is most closely related to Ig and Tcr V regions. Although the amino-terminal domain of L3T4 is the portion of the molecule that is most similar to V-regions, L3T4 is one of the polydomain members of the Ig gene superfamily.

Studies of the expression of L3T4 mRNA in various tissues led to the surprising finding that this gene is transcribed not only in T lymphoid cells, but also in brain. The predominant form of L3T4 mRNA expressed in brain is foreshortened as compared to that in T lineage cells, and it is most likely the product of a distinct transcripitional start site. If translated, the protein encoded by this brain transcript would be 217 amino acids in length and would lack the signal peptide and the amino-terminal 214 amino acids of the mature protein. It is not known whether a stable protein product is synthesized from this mRNA or what its function might be. However, these findings place L3T4 in an intriguing class of Ig gene superfamily members characterized by coexpression in the immune system and the nervous system.

ACKNOWLEDGMENTS

We would like to thank Kim Brown for help in preparing the manuscript. This work was supported by NIH grants GM34991 and CA46507. J. R. P. was the recipient of a John A. and George L. Hartford faculty fellowship award.

REFERENCES

Amzel, L. M. & Poljak, R. J. (1979) Three-dimensional structure of immunoglobulins. *Ann. Rev. Biochem.* **48**, 961.

Ceredig, R., Dialynas, D. P., Fitch, F. W. & MacDonald, H. R. (1983) Precursors of T cell growth factor producing cells in the thymus: Ontogeny, frequency, and quantitative recovery in a subpopulation of phenotypically mature thymocytes defined by monoclonal antibody GK-1.5. *J. Exp. Med.* **158**, 1654.

Clark, S. J., Jefferies, W. A., Barclay, A. N., Gagnon, J. & Williams, A. F. (1987) Peptide and nucleotide sequences of rat CD4 (W3/25) antigen: Evidence for derivation from a structure with four immunoglobulin-related domains. *Proc. Natl. Acad. Sci. USA* **84**, 1649.

Classon, B. J., Tsagaratos, J., McKenzie, I. F. C. & Walker, I. D. (1986) Partial primary structure of the T4 antigens of mouse and sheep: Assignment of intrachain disulfide bonds. *Proc. Natl. Acad. Sci. USA* **83**, 4499.

Cohen, F. E., Novotny, J., Sternberg, M. J. E., Campbell, D. G. & Williams, A. F. (1981) Analysis of the structural similarities between brain Thy-1 antigen and immunoglobulin domains. *Biochem. J.* **195**, 31.

Crocker, P. R., Jefferies, W. A., Clark, S. J., Chung, L. P. & Gordon, S. (1987) Species heterogeneity in macrophage expression of the CD4 antigen. *J. Exp. Med.* **166**, 613.

Dalgleish, A. G., Beverly, P. C. L., Clapham, P. R., Crawford, D. H., Greaves, M. R. & Weiss, R. A. (1984) The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. *Nature* **312**, 763.

Dembic, Z., Hass, W., Zamoyska, R., Parnes, J., Steinmetz, M. & von Boehmer, H. (1987) Transfection of the CD8 gene enhances T-cell recognition. *Nature* **326**, 510.

Dialynas, D. P., Wilde, D. B., Marrack, P., Pierres, A., Wall, K. A., Havran, W., Otten, G., Loken, M. R., Pierres, M., Kappler, J. & Fitch, F. W. (1983) Characterization of the murine antigenic determinant, designated L3T4a, recognized by a monoclonal antibody GK1.5: Expression of L3T4a by functional T cell clones appears to correlate primarily with class II MHC antigen-reactivity. *Immunological Rev.* **74**, 29.

Epstein, L. G., Sharer, L. P., Cho, E.-S., Myenhofer, M., Navia, B. A. & Price, R. W. (1985) HTLV-III/LAV-like retrovirus particles in the brains of patients with AIDS encephalopathy. *AIDS Res.* **1**, 447.

Field, E. H., Tourvieille, B., D'Eustachio, P. & Parnes, J. R. (1987) The gene encoding the mouse T cell differentiation antigen L3T4 is located on chromosome 6. *J. Immunol.* **138**, 1968.

Funke, I., Hahn, A., Rieber, E. P., Weiss, E. & Riethmuller, G. (1987) The cellular receptor (CD4) of the human immunodeficiency virus is expressed on neurons and glial cells in human brain. *J. Exp. Med.* **165**, 1230.

Gabert, J., Langlet, C., Zamoyska, R., Parnes, J. R., Schmitt-Verhulst, A.-M. & Malissen, B. (1987) Reconstitution of MHC-class I specificity by T-cell receptor and Lyt-2 gene transfer. *Cell* **50**, 545.

Gorman, S. D., Tourvieille, B. & Parnes, J. R. (1987) Structure of the mouse CD4 gene and an unusual transcript in brain. *Proc. Natl. Acad. Sci. USA* **84**, 7644.

Gottlieb, P. D. (1974) Genetic correlation of a mouse light chain variable region marker with a thymocyte surface antigen. *J. Exp. Med.* **140**, 1432.

Hill, R. L., Delaney, R., Fellow, R. E., Jr., & Lebowitz, H. E. (1966) The evolutionary origins of the immunoglobulins. *Proc. Nat. Acad. Sci.* **80**, 2995.

Ho, D. D., Rota, T. R., Schooley, R. T., Kaplan, J. C., Allan, J. D., Groopman, J. E., Resnick, L., Felsenstein, D., Andrews, C. A. & Hirsch, M. S. (1985) Isolation of HTLV-III from cerebrospinal fluid and neural tissues of patients with neurological syndromes related to the acquired immunodeficiency syndrome. *N. Engl. J. Med.* **313**, 1498.

Hollander, N., Pillemer, E. & Weissman, I. L. (1980) Blocking effect of Lyt-2 antibodies on T cell functions. *J. Exp. Med.* **152**, 674.

Hunig, T. (1984) Monoclonal anti-Lyt-2.2 antibody blocks lectin-dependent cellular cytotoxicity of H-2 negative target cells. *J. Exp. Med.* **159**, 551.

Hunkapiller, T. & Hood, L. (1986) Growing immunoglobulin gene superfamily. *Nature* **323**, 15.

Hunkapiller, T. & Hood, L. (1987) Diversity of the immunoglobulin gene superfamily. *Adv. Immun.* (in press).

Jefferies, W. A., Green, J. R. & Williams, A. F. (1985) Authentic T helper CD4 (W3/25) antigen on rat peritoneal macrophages. *J. Exp. Med.* **162**, 117.

Johnson, P. & Williams, A. F. (1986) Striking similarities between antigen receptor J pieces and the sequence of the second chain of the murine CD8 antigen. *Nature* **323**, 74.

Johnson, R. T. & McArthur, J. D. (1986) AIDS and the brain. *TINS* **9**, 91.

Klatzman, D., Champagne, E., Chamaret, S., Gruest, J., Guetard, D., Hercend, T., Gluckman, J.-C. & Montagnier, L. (1984) T-lymphocyte T4 molecule behaves as the receptor for human retrovirus LAV. *Nature* **312**, 767.

Koenig, S., Gendelman, H. E., Orenstein, J. M., Dal Canto, M. C., Pezeshkpour, G. H., Yungbluth, M., Janotta, F., Aksamit, P., Martin, M. A. & Fauci, A. S. (1986) Detection of AIDS virus in macrophages in brain tissue from AIDS patients with encephalopathy. *Science* **233**, 1089.

Levy, J. A., Hollander, H., Shimabukoro, J., Mills, J. & Kaminsky, L. (1985) Isolation of AIDS-associated retrovirus from cerebrospinal fluid and brain of patients with neurological symptoms. *Lancet* **II**, 586.

Littman, D. R. & Gettner, S. N. (1987) Unusual intron in the immunoglobulin domain of the newly isolated murine CD4 (L3T4) gene. *Nature* **325**, 453.

MacDonald, H. R., Glasebrook, A. L., Bron, C., Kelso, A. & Cerottini, J.-C. (1982) Clonal heterogeneity in the functional requirement for Lyt-2/3 molecules on cytolytic T lymphocytes (CTL): Possible implications for the affinity of CTL antigen receptors. *Immunological Rev.* **68**, 89.

MacDonald, H. R., Thierness, N. & Cerottini, J.-C. (1981) Inhibition of T cell-mediated cytolysis by monoclonal antibodies directed against Lyt-2: Heterogeneity of inhibition at the clonal level. *J. Immunol.* **126**, 1671.

Maddon, P. J., Dalgleish, A. G., McDougal, J. S., Clapham, P. R., Weiss, R. A. & Axel, R. (1986) The T4 gene encodes the AIDS virus receptor and is expressed in the immune system and the brain. *Cell* **47**, 333.

Maddon, P. J., Littman, D. R., Godfrey, M., Maddon, D. E., Chess, L. & Axel, R. (1985) The isolation and nucleotide sequence of a cDNA encoding the T cell surface protein T4: A new member of the immunoglobulin gene family. *Cell* **42**, 93.

Marrack, P., Endres, R., Shimonkevitz, R., Zlotnik, A., Dialynas, D., Fitch, F. & Kappler, J. (1983) The major histocompatibility complex-restricted antigen receptor on T cells. II. Role of the L3T4 gene product. *J. Exp. Med.* **158**, 1077.

Mathieson, B. J. & Fowlkes, B. J. (1984) Cell surface antigen expression on thymocytes:

Development and phenotypic differentiation of intrathymic subsets. *Immunological Rev.* **82**, 141.

Matsunaga, T. & Mori, N. (1987) The origin of the immune system. *Scand. J. Immunol.* **25**, 485.

McDougal, J. S., Kennedy, M. S., Sligh, J. M., Cort, S. P., Mawle, A. & Nicholson, J. K. A. (1986) Binding of HTLV-III/LAV to T4+ T cells by a complex of the 110k viral protein and the T4 molecule. *Science* **231**, 382.

Moscicki, R. A., Amento, E. P., Krane, S. M., Kurnick, J. T. & Colvin, R. B. (1983) Modulation of surface antigens of a human monocyte cell line, U937, during incubation with T-lymphocyte-conditioned medium: Detection of T4 antigen and its presence on normal blood monocytes. *J. Immunol.* **131**, 743.

Nakauchi, H., Nolan, G. P., Hsu, C., Huang, H. S., Kavathas, P. & Herzenberg, L. A. (1985) Molecular cloning of Lyt-2, a membrane glycoprotein marking a subset of mouse T lymphocytes: molecular homology to its human counterpart, Leu-2/T8, and to immunoglobulin variable regions. *Proc. Natl. Acad. Sci. USA* **82**, 5126.

Owens, G. C., Edelman, G. M. & Cunningham, G. A. (1987) Organization of the neural cell adhesion molecule (N-CAM) gene: Alternative exon usage as the basis for different membrane-associated domains. *Proc. Nat. Acad. Sci.* **84**, 284.

Peault, B., Chen, C.-L. H., Cooper, M. D., Barbu, M., Lipinski, M. & le Douarin, N. M. (1987) Phylogenetically conserved antigen on nerve cells and lymphocytes resembles myelin-associated glycoprotein. *Proc. Nat. Acad. Sci.* **84**, 814.

Pert, C. B., Hill, J. M., Ruff, M. R., Bermon, R. M., Robey, W. G., Arthur, L. O., Russetti, F. W. & Farrar, W. L. (1986) Octapeptides deduced from the neuropeptide receptor-like pattern of antigen T4 in brain potently inhibit human immunodeficiency virus receptor binding and T-cell affinity. *Proc. Natl. Acad. Sci. USA* **83**, 9254.

Sarmiento, M., Glasebrook, A. L. & Fitch, F. W. (1980) IgG or IgM monoclonal antibodies reactive with different determinants on the molecular complex bearing Lyt-2 antigen block T cell-mediated cytolysis in the absence of complement. *J. Immunol.* **125**, 2665.

Scollay, R., Bartlett, P. & Shortman, K. (1984) T cell development in the adult murine thymus: changes in the expression of the surface antigens Ly2, L3T4 and B2A2 during development from early precursor cells to emigrants. *Immunological Rev.* **82**, 79.

Sharer, L. R., Epstein, L. G., Cho, E.-S., Joshi, V. V., Meyenhofer, M. F., Rankin, L. F. & Petito, C. K. (1986) Pathological features of AIDS encephalophaty in children: evidence for LAV HTLV-III infection of brain. *Human Path.* **17**, 271.

Shaw, G. H., Harper, M. E., Hahn, B. H., Epstein, L. G., Gajdusek, D. C., Price, R. W., Navia, B. A., Petito, C. K., O'Hara, C. J., Groopman, J. E., Cho, E. S., Oleske, J. M., Wong-Staal, F. & Gallo, R. C. (1985) HTLV-III infection in brains of children and adults with AIDS encephalopathy. *Science* **227**, 177.

Snider, W. D., Simpson, D. M., Nielsen, S., Gold, J. W. N., Metroka, C. E. & Posner, J. B. (1983) Neurological complications of acquired immune deficiency syndrome: Analysis of 50 patients. *Ann. Neurol.* **14**, 403.

Swain, S. L. (1983) T cell subsets and the recognition of MHC class. *Immunological Rev.* **74**, 129.

Thompson, J. A., Pande, H., Paxton, R. J., Shively, L., Padma, A., Simmer, R. L., Todd, C. W., Riggs, A. D. & Shively, J. E. (1987) Molecular cloning of a gene belonging to the carcinoembryonic antigen gene family and discussion of a domain model. *Proc. Natl. Acad. Sci.* **84**, 2965.

Tite, J. P., Sloan, A. & Janeway, C. A., Jr. (1986) The role of L3T4 in T cell activation: L3T4 may be both an Ia-binding protein and a receptor than transduces a negative signal. *J. Mol. Cell. Immunol.* **2**, 179.

Tse, A. G. D., Barclay, A. N., Watts, A. & Williams, A. F. (1985) A glycophospholipid tail

at the carboxyl-terminus of the Thy-1 glycoprotein of neurons and thymocytes. *Science* **230**, 1003.

Wassmer, P., Chan, C., Logdberg, L. & Shevach, E. M. (1985) Role of the L3T4 antigen in T cell activation. II. Inhibition of T cell activation by monoclonal anti-L3T4 antibodies in the absence of accessory cells. *J. Immunol.* **135**, 2237.

Wiley, C. A., Schrier, R. D., Nelson, J. A., Lampert, P. W. & Oldstone, M. B. A. (1986) Cellular localization of human immunodeficiency virus infection within the brains of acquired immune deficiency syndrome patients. *Proc. Natl. Acad. Sci. USA* **83**, 7089.

Williams, A. (1982) Surface molecules and cell interactions. *J. Theor. Biol.* **98**, 12.

Wood, G. S., Warner, N. L. & Warnke, R. A. (1983) Anti-Leu-3/T4 antibodies react with cells of monocyte-macrophage and Langerhans lineage. *J. Immunol.* **131**, 212.

Zimmermann, W., Ortlieb, B., Friedich, R. & von Kleist, S. (1987) Isolation and characterization of cDNA clones encoding the human carcinoembryonic antigen reveal a highly conserved repeating structure. *Proc. Natl. Acad. Sci.* **84**, 2960.

Implications of the Diversity of the Immunoglobulin Gene Superfamily

T. HUNKAPILLER, J. GOVERMAN, B.F. KOOP, AND L. HOOD
Division of Biology, California Institute of Technology, Pasadena, California 91125

The vertebrate immune response consists of a complex set of cellular and serologic reactions that provide protection from infectious foreign agents and abnormal or damaged cells. Responses to these insults are mediated by a complex array of immune elements that recognize distinct macromolecular structural patterns (antigens). Characterization of these elements over the past two decades has shown that many share a common evolutionary precursor, a gene element encoding the immunoglobulin homology unit (Cunningham et al. 1973; Strominger et al. 1980; Williams 1984; Hood et al. 1985). Recently, a large number of molecules, with no immunologic function and mostly expressed on a cell surface, have also been shown to share this same ancestral element (Williams 1987; Hunkapiller and Hood 1989). Therefore, the genes encoding these related molecules are defined as belonging to the immunoglobulin gene superfamily. Together, these genes encode an amazingly diverse array of functions from immune recognition to cellular adhesion, reflecting the versatility of the shared immunoglobulin homology unit (Fig. 1).

The immunoglobulin gene superfamily includes many unique genes as well as numerous multigene families. The potential to duplicate, delete, and reorganize informational units at all levels in the hierarchal organization of such a superfamily—nucleotides, exons, genes, and even multigene families—increases the possibilities for rapid evolutionary change of complex phenotypic characters. This potential has profoundly affected views about the driving forces of evolution, emphasizing the possibility of rapid acquisition of novel phenotypic traits. Moreover, the fact that the immunoglobulin homology unit is employed in many different types of systems argues that there is an underlying unity of shared strategies for molecular recognition at the cell surface. Thus, the immunoglobulin gene superfamily is a common thread weaving through the complex patterns of eukaryotic biology and evolution.

In this paper, we first define the homology unit and describe the features that lend particular advantage to its use in such a wide range of biological contexts. We illustrate the diversity and combinatorial advantages of the homology unit through a consideration of the strategies of molecular recognition in the immune system. The ability of the homology unit to accommodate tremendous functional flexibility is discussed primarily in the context of nonimmune receptor molecules. We then discuss the evolution of the immunoglobulin gene superfamily and the implications it has for eukaryotic biology.

Immunoglobulin Homology Unit

Molecules are defined as belonging to the immunoglobulin superfamily by the presence of one or more regions homologous to the basic structural unit of immunoglobulins, the immunoglobulin homology unit (Hill et al. 1966) (Fig. 2). These regions are usually between 70 and 110 residues in length and are characterized by a series of nonparallel β strands that generate a compact sandwich of two β sheets (see Amzel and Poljak 1979). This structure is stabilized by a small series of conserved residues, particularly, two virtually invariant cysteine residues that generate a signature disulfide bond holding the faces of the β sheet sandwich together. The loop sequences connecting the β strands are less significant for the basic homology unit structure and therefore, can accommodate extensive variability.

Three distinct types of immunoglobulin homology units have been identified (Williams 1987; Hunkapiller and Hood 1989). These are compared in Figure 3 and are characterized by type-specific amino acid residues and the presence or absence of particular β strands and loop sequences between the two cysteine residues forming the characteristic disulfide bridge. These homology units include the V and C homology units characteristic of the immune-antigen receptors (discussed later), both defined from crystallographic analysis (Fig. 2). The V unit has an extra β-strand pair and loop sequence relative to C units. The extra loop means the V motif can accommodate even greater variability. We have denoted a third unique homology unit type, H. This motif appears to have a more C-like structure, lacking the same β-strand pair and loop, but it has a slightly more V-like sequence. Also, H homology units are the primary motif of various molecules that seem to predate in evolution the molecules of the immune receptors. Hence, V and C units appear to be the more specialized versions, whereas H units more nearly represent the primordial homology unit motif.

The number and functional diversity of molecules incorporating the immunoglobulin homology unit is proof that it has been a tremendously successful protein motif throughout evolution. Four particular factors have contributed to this success (Hunkapiller and

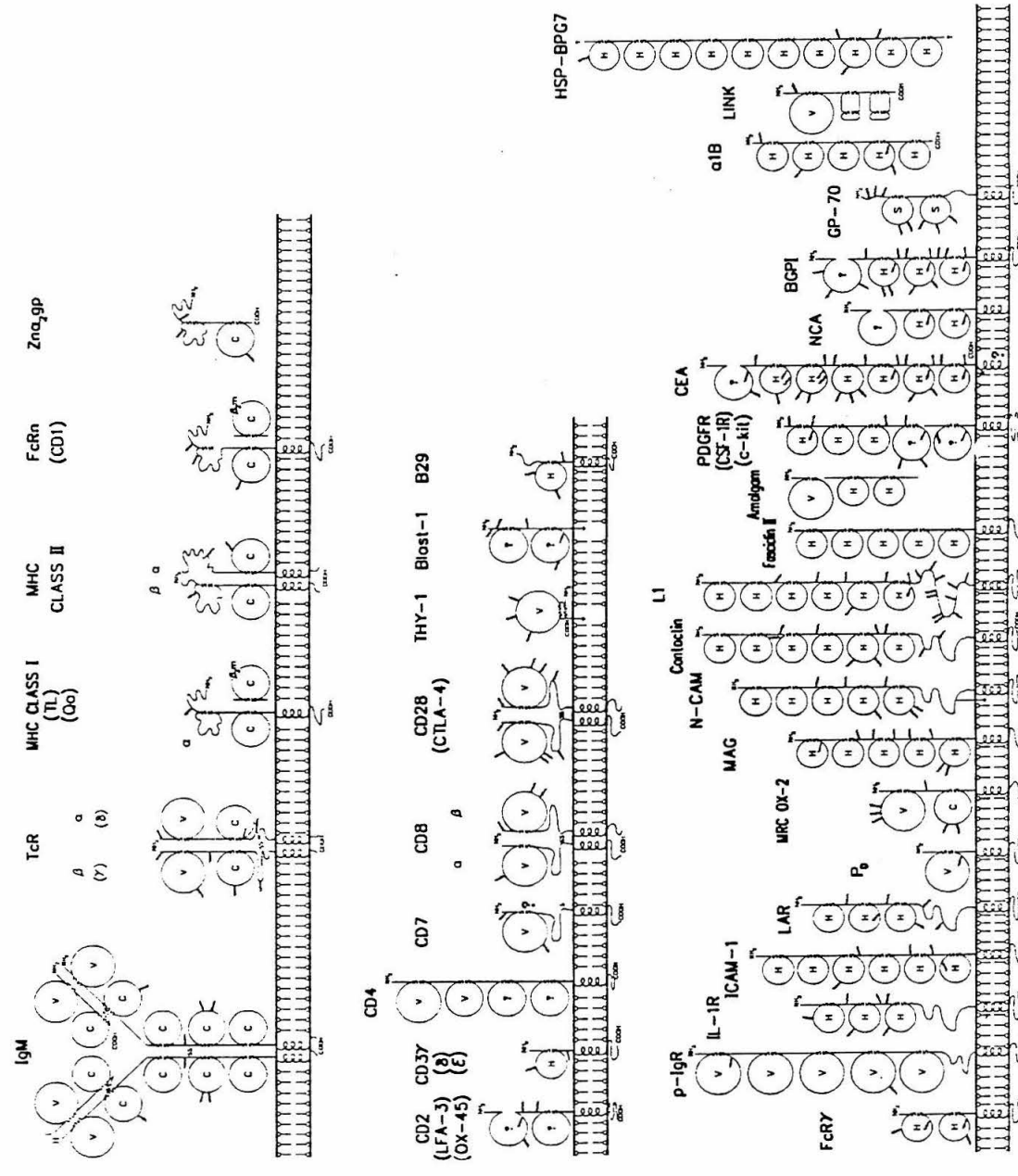


Figure 1. (*See facing page for legend.*)

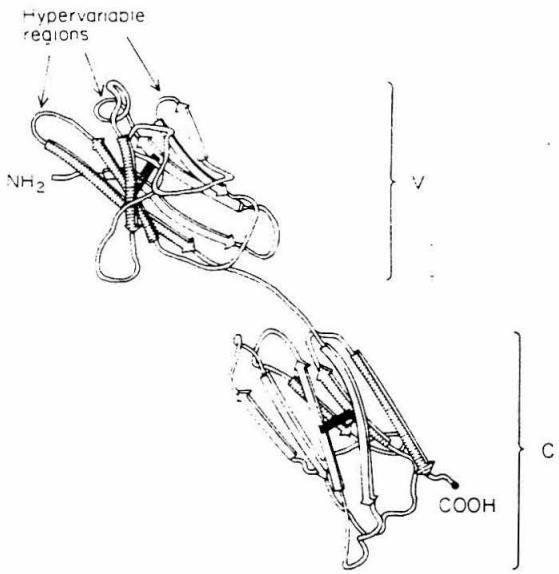


Figure 2. Tertiary organization of V and C homology units (after Edmundson et al. 1975). This illustrates the tertiary structure of an immunoglobulin light chain. The β strands and their orientations are shown as flat arrows. Opposite faces of the sandwich structure are indicated by either blank or hatched β strands. The disulfide bond is indicated by a solid bar. (Reprinted, with permission, from Hunkapiller and Hood 1989)

Hood 1989). First, apart from the very few conserved residues involved in direct inter- and intrachain interactions, the primary structure of the homology unit can vary dramatically and still generate the same tertiary structure. The loops that connect the β-pleated sheet strands are particularly variable. Thus, because the homology unit is capable of accommodating enormous diversity of primary sequence, it can be the structural foundation for an equally enormous array of highly specific interactions. Second, the tertiary structure of the homology unit is such that homology units tend to interact with one another. This impacts both the development of the homo- and/or heterotypic dimers that are the basis of many receptor structures as well as the protein-protein interactions between different or the same immunoglobulin superfamily members. Combinatorial associations between different protein chains mediated by homology unit interactions increase both the evolutionary and somatic diversification potentials of the gene systems involved. Such interactions should favor the establishment of new functional associations between existing members of the immunoglobulin gene superfamily. The third feature that contributes to the evolutionary advantage of the homology unit is its striking stability. It must be an effective cell-surface display structure for molecular recognition, and, as such, it must be resistant to proteolytic digestion by the enzymes that bathe these external receptors. Thus, the need for a compact, globular structure resistant to proteolysis may have been the initial driving force in the evolution of the homology unit. Finally, most homology units are encoded by discrete exons. The 3′ end of each exon ends with the first base of the next codon, whereas each 5′ end has the last two bases. These conserved splicing rules ensure correct, in-frame translation of any number of tandemly encoded units regardless of their order. These exons, therefore, serve as a basic unit of evolutionary diversification.

Molecular Recognition in the Immune System

There are four distinct classes of immune receptors, molecules that interact directly with antigen. The immunoglobulin homology unit was first recognized during the study of antibodies, the immunoreceptors of B cells (Hill et al. 1966). Antibodies can be either fixed to the cell surface or expressed humorally; that is, secreted into the blood to carry out their functions at a distance from the cell of origin. T cells, on the other hand, synthesize only a surface-bound antigen receptor. There are also two classes of receptors encoded within the major histocompatibility complex (MHC). Class I MHC molecules are present in varying amounts on virtually all somatic cells. Class II MHC molecules are expressed only on particular antigen-presenting cells (discussed later) such as macrophages and B cells. The multiple roles played by the homology units in these molecules dramatically reflect the versatility of the basic motif.

The basic structure of antibodies is a tetramer composed of identical heterodimers of a light and a heavy chain (Fig. 1) (see Klein 1982). The antibody molecule is divided into variable (V) domains that are responsible for pattern recognition and constant (C) domains that carry out various effector functions. The V domain is constructed of one V homology unit from both the light (V_L) and heavy (V_H) chains. The C region of light chains (C_L) likewise has only a single C homology unit, whereas the C region of heavy chains (C_H) has two to four C homology units, depending on the type of heavy

Figure 1. Schematic diagram of the members of the immunoglobulin gene superfamily (IgGSF). Disulfide bonds are represented by (S–S). Homology units are indicated as loops labeled V, C, or H (see text). Loops of uncertain relationship to homology units are labeled with a question mark. Different-sized loops illustrate the relative differences in length between the conserved disulfide bond of the labeled types. Membrane-spanning peptides are shown as simple helices. Glycophospholipid linkage to the membrane is represented as an arrow. Intracytoplasmic regions are drawn with wavy lines that indicate their relative lengths. Extra- and intracellular orientations are indicated by NH_2 and COOH labels on the protein chains, respectively. Possible asparagine-linked carbohydrates are shown as jagged lines extending from the protein chains. Note that these sites are not necessarily conserved between alleles or across species, but are representative of at least one known example of the labeled protein. Related IgGSF members are illustrated with a single structure, as indicated by the name labels above each structure. For a description of the molecules and their references, see Hunkapiller and Hood (1989).

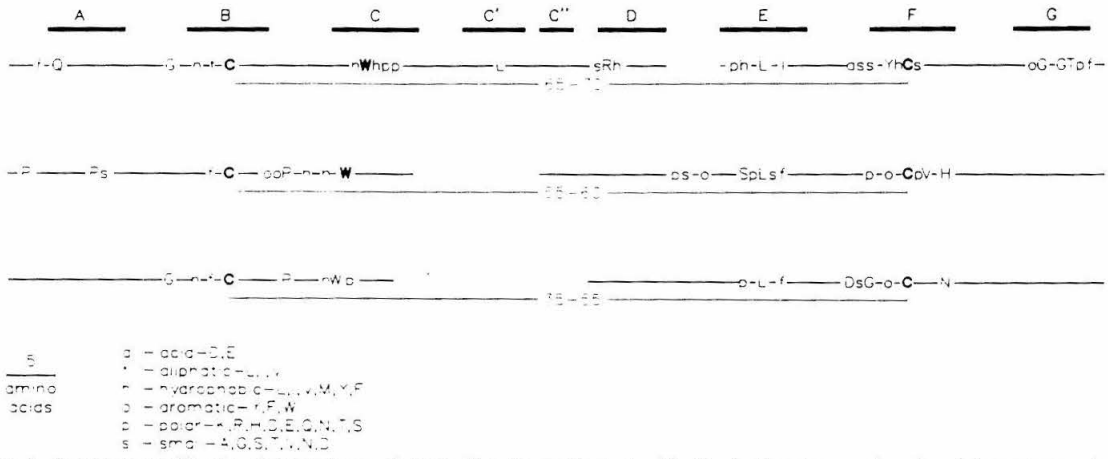


Figure 3. Sequence motifs of each homology unit type. This figure illustrates idealized aligned examples of each homology unit type. Representative amino acid residues of each type are shown in their relative position along the polypeptides. Uppercase letters are single-letter amino acid code. Lowercase letters represent a selection of amino acids with similar functional or physical properties. The key to this code is shown in the lower left. Relative gaps are introduced into each chain to align residues conserved between types. The more nearly invariant cysteine and tryptophan residues are highlighted. The distances between the conserved cysteine residues are shown beneath each chain. The approximate locations of the β strands are illustrated with heavy bars above the sequence representations. Note the two β strands (C' and C'') not found in the C and H domains. It is also clear that β strand D may not be found in the shorter H domain sequences. (Reprinted, with permission, from Hunkapiller and Hood 1989).

chain. As with the V regions, pairs of C-region homology units interact to form discrete structural domains. There are several different C_H genes, each of which defines a different class or isotype of immunoglobulin. Different effector functions and interactions are mediated by different classes of antibodies and facilitate the destruction or elimination of antigen or antigen-associated pathogens.

The T-cell antigen receptor is a simple heterodimer with one V domain involved in antigen recognition and one C domain that interacts with a membrane-bound protein complex, CD3, presumably involved in signal transduction (Fig. 1) (see Davis 1985; Kronenberg et al. 1986). T cells responsible for general antigen recognition use α and β chains (see Hannum et al. 1984; Meuer et al. 1984). A small subset of T cells employs γ and δ chains (Bank et al. 1986; Brenner et al. 1986; Ioannidis et al. 1987). The function of the γ/δ cells is unclear, but some form of interaction with particular antigens appears likely (e.g., Matis et al. 1987). Unlike B cells, T cells do not express different isotypes. Rather, there are subsets of T cells that use common receptor recognition gene elements, but have different functional responses upon activation, presumably due to the expression of other, subset-specific gene products. This is consequent with the lack of direct effector function of the T-cell receptor itself. Cytotoxic T cells (T_c) destroy altered or damaged cells, often as the result of viral infections or neoplastic transformation. Helper T cells (T_H) facilitate the differentiation and activation of B cells and other T cells and, accordingly, play a central role in the regulatory and developmental processes of the immune system.

The primary sequences responsible for antigenic interaction and specificity in immunoglobulins are the four relatively hypervariable loop sequences between the β strands of the V homology unit faces (see Amzel and Poljak 1979). Homologous sequences are presumed to function the same for T-cell receptors (Goverman et al. 1986; Novotný et al. 1986). These hypervariable regions make up the antigen-binding pocket or cleft (Wu and Kabat 1970; Capra and Kehoe 1974) and, consequently, play a critical role in determining the antigen-specific repertoire. The ability of V regions to accommodate diversity is illustrated by the fact that even V gene regions of the same gene family can have less than 20% amino acid similarity.

Molecular recognition in the T and B cell systems is mediated by distinctly different mechanisms. Antibodies recognize the native three-dimensional structure of macromolecules by virtue of their molecular complementarity to the foreign structural pattern. T cells expressing α/β receptors, on the other hand, recognize only peptide fragments of antigens when they are bound by a receptor encoded in the MHC expressed on an antigen-presenting cell (see Schwartz 1984). As a result, T-cell activation requires the formation of a trimolecular complex, the T-cell receptor recognizing portions of both the antigenic peptide and the MHC molecule. Thus, the complex ligand of the T-cell receptor bears no necessary structural similarity to the native antigenic protein. This requirement for direct cell-cell interaction to activate T cells is perhaps the primary operational distinction between T and B cells and is critical to the regulatory role played by T cells in the immune response.

A major aspect of this regulatory role of T cells is establishing and maintaining tolerance to self-antigens. This is accomplished in large part during T-cell development in the thymus as T cells that react strongly to self-molecules are clonally deleted (see Sprent and Webb 1987). A subset of the cells remaining become

responsible for regulation of B-cell activation. Communication between B and T cells is established through the role of the B cell as an antigen-presenting cell. That is, B cells bind native antigen with surface-bound immunoglobulin, process the protein, then present denatured fragments via class II MHC molecules to T cells. Thus, T cells are the major arbiters of whether or not B cells proliferate and differentiate to produce quantities of antibodies after they have interacted with their complementary antigen.

There are two unique features shared by both B and T cells. First, through numerous genomic and somatic diversification strategies (discussed later), each set of cells can express an enormous diversity of immune receptors. Second, each individual B and T cell generally expresses only a single clonotype of receptor molecule, a process referred to as allelic exclusion (Luzzati et al. 1973). This clonal expression provides a means whereby each B or T cell can express a unique receptor type with a unique specificity. Therefore, any particular antigen will selectively amplify only the appropriate subset of T or B cells that have complementary receptors. This antigen-specific expansion is referred to as clonal selection. Thus, the primary consequence of allelic exclusion is the ability to regulate finely the immune response to particular antigens in the context of the universe of all possible antigens.

MHC molecules are not classic receptors, in that their association with their ligand peptide does not occur on the cell surface. Rather, the peptides are products of intracytoplasmic enzymatic processing and become associated with MHC molecules before cell-surface expression (see Claverie and Kourilsky 1987). Class I MHC molecules present bound peptide primarily to T_c cells, whereas class II molecules present peptides generally to T_H cells (see Schwartz 1984). The class I molecule is a heterodimer composed of the highly polymorphic class I polypeptide and the invariant, non-MHC-encoded β_2-microglobulin. The class II molecule is a heterodimer composed of two highly polymorphic chains, α and β. All four chains contain one C homology unit (Fig.1) (see Hood et al. 1983). In the case of each pair, the association is mediated by noncovalent interactions of the C homology units. All but β_2-microglobulin also contain other, non-homology unit regions that are responsible for peptide binding and antigen presentation (see Germain and Malissen 1986). Unlike B- and T-cell receptors, each individual expresses only a limited number of different MHC molecules. Therefore, the peptide-binding sites must be very degenerate in order to accommodate the very large number of potential, different peptide fragments. Also, because these genes are not allelically excluded, all the different MHC molecules expressed in an individual are co-expressed on the same cells.

Diversity of the Immune Receptors

The genes encoding antibodies are divided into three unlinked gene families; two encode light chains (λ and κ), and the third encodes heavy chains (Fig. 4) (see Honjo 1983). Likewise, the genes encoding the T-cell receptors are divided into three gene families; two encode the α and β chains, and a third encodes the γ chains (Fig. 4) (Davis 1985; Hayday et al. 1985; Kronenberg et al. 1986). The δ genes are contained within the α-chain locus (Chien et al. 1987). The variable regions of antibodies and T-cell receptors are encoded by two or three gene segments that are separate in the germ line: variable (V) and joining (J) for light, α and γ genes and V, diversity (D), and J for heavy, β and δ genes (Fig. 4). The V, D, and J or V and J elements are brought together through somatic DNA rearrangement events to produce complete, expressible V genes.

A large variety of somatic and evolutionary strategies for diversification are shared by T- and B-cell receptors. The basis of the somatic strategies is the rearrangement process. In mammals, this joining is able to unite virtually any VJ, DJ, or VD gene segments within a family, generating by combinatorial means much greater diversity potential than just the number of germ-line elements would support (Table 1). Also, the joining process itself is imprecise and generates enormous additional molecular diversity at the junctional

Table 1. V Gene Diversity

	α	β	Heavy	κ
V gene segments	100	30	500	200
(subfamilies)	(14)	(17)	(8)	(5)
D gene segments	—	2	15	—
J gene segments	50	6 + 6	4	4
Ds in 3 reading frames	n.a.	++	+	n.a.
N-region sequence	++	++	++	−
Junctional diversity	+++	+++	+++	+
Somatic hypermutation	−	−	+	+
Alternate joining order	−	+?	+?	−
Combinatorial joining	V × J 100 × 50	V × D × J (30 × 3 × 12) +(30 × 3 × 6)	V × D × J 500 × 15 × 4	V × J 200 × 4
Total	5000	1620	3×10^4	800
Combinatorial association	× 8.1×10^6		× 2.4×10^7	

n.a. indicates not applicable

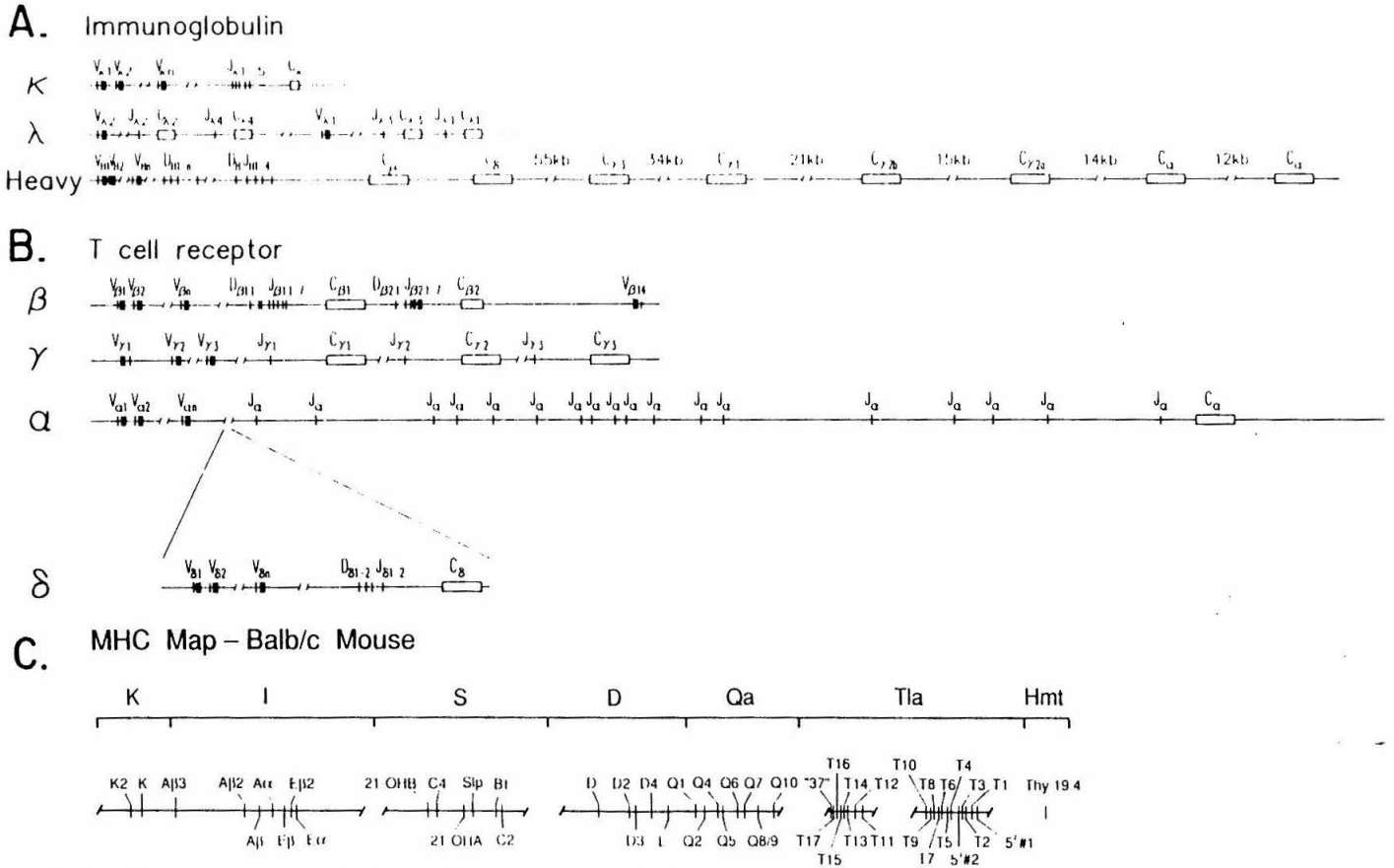


Figure 4. Organization of the murine immunoglobulin (*A*), T-cell receptor (*B*), and MHC (*C*) gene families. Open boxes in *A* and *B* are C-region genes, and filled boxes are V-region gene segments. Vertical lines illustrate D and J gene segments. Numbers between C-region genes in the heavy-chain family are distances in kilobases. Double slashes without numbers indicate unknown distances. Human gene families are organized similarly. The δ gene family is shown as an insert to the α gene family. The MHC map indicates class I regions (e.g., K, I, S, D, Qa, Tla, Hmt), class II regions (A and E), and the complement associated S-region class III genes. Individual genes are indicated below.

regions through various mechanisms (see Tonegawa 1983; Hunkapiller and Hood 1989). These mechanisms lead to such extreme variability in the junction regions that a window of variability is generated between the V and J elements within which virtually any amino acid may be represented at any position (Elliot et al. 1988; Hunkapiller and Hood 1989). Depending on the gene family, these windows range in size from 3 to 15 residues. This junctional window comprises the third hypervariable loop of the antigen-binding site. Accordingly, variability here has a critical impact on the functional diversity of antibody and T-cell receptor repertoires.

A further mechanism for diversification is a consequence of the fact that essentially any light chain may associate with any heavy chain (or any α with any β). Thus, a final level of molecular diversity is generated through combinatorial association at the polypeptide chain level (Table 1). The fact that the gene families are not linked increases the potential for diversity at the level of populations even further by dissociating the allelism of the families. This provides the possibility, within a population over time, for any V_H allele to combine with any V_L allele.

B cells have an additional mechanism for antigen-specific diversification called somatic hypermutation (Kim et al. 1981; Gearhart and Bogenhagen 1983). This process scatters random single-base substitutions throughout the rearranged V gene, generating variant antibodies with altered affinity for the antigen. Thus, the presence of antigen can selectively expand those B cells producing variant antibodies with an increased affinity for the stimulating antigen. Seen in normal immune responses, this process is called affinity maturation (Gearhart et al. 1981; Clarke et al. 1985; Sablitzky et al. 1985). T cells do not undergo somatic hypermutation. Presumably, this inability to diversify further after leaving the thymus is linked to their role as regulators of the immune response (Barth et al. 1985; Honjo and Habu 1985; Eisen 1986). Clonal deletion of self-reactive T cells occurs during thymic education. If T cells were to mutate randomly their receptors after leaving the thymus, there would be no check on the accidental development of harmful, self-reactive T cells. B cells can take this risk because T cells can still maintain activation control of B cells with receptors for self-antigens.

B cells also employ two other novel strategies for diversifying the immune response potential. However, these do not affect the possible range of antigen specificities, but rather the functional, temporal, and spatial distributions of the possible receptor specificities. Both are related to the lack of any requirement for cell-cell interactions. First, alternative patterns of RNA splicing of heavy-chain transcripts can generate molecules with or without hydrophobic transmembrane regions (Early et al. 1980). Thus, antibodies with the same antigen specificity can be alternatively bound to the cell surface to serve as receptors for the differentiation of B cells or secreted into the blood to serve as end-stage effector molecules. Second, the assembled V_H chain gene can rearrange into juxtaposition with different C_H genes during a process called class switching (Davis et al. 1980). In this manner, the same molecular recognition domain can be expressed with very different kinds of effector or tissue-targeting domains. Thus, the great variability in ligand specificity provided by the V homology unit can be expressed in the context of the functional and structural variability accommodated by the C homology unit.

With the tremendous potential for somatic diversification, the repertoire of unique T- and B-cell receptors generated in one individual is, in theory, limited only by the number of T- and B-progenitor cells that have been generated over the lifetime of that individual. When somatic hypermutation is considered, the potential for variety in antibodies begins to approach the number of actual B cells. This variation is uniquely accommodated by the V homology unit, allowing essentially the somatic tailoring of antigen-receptor repertoires to the current antigenic environment.

Diversity of the MHC genes is based solely on evolutionary mechanisms and the rules of Mendelian genetics. The class I molecules that present antigen in the mouse are encoded by only two or three genes (with a larger number of nonclassic MHC class I genes of unknown function denoted Q, T, and Hmt), and the class II molecules are encoded by two to four α and three to five β genes (Fig. 4) (see Hood et al. 1983; Mengle-Gaw and McDevitt 1985). Those portions of the genes that encode the peptide-binding sites constitute perhaps the most polymorphic non-viral protein-coding sequences known (see Klein and Figueroa 1981). These variations are generated by point mutations as well as recombinational and gene conversion events involving the genes of the non-antigen-binding MHC molecules (see Klein and Figueroa 1986). Thus, although the diversity within populations of expressed collections of MHC molecules can be very large, individuals can only express a relatively small degree of this diversity. The extreme allelism maintained in populations ensures the widest range of variation of peptide-presenting molecules possible for an individual (Hunkapiller and Hood 1989). The class II genes can also generate diversity by combinatorial association of their discrete polypeptide chains. However, the close linkage between the α and β genes tends to reduce this potential within a population.

The diversification strategies of the B- and T-cell receptors, when compared to those of the MHC molecules, reflect basically the different requirements of target and effector cells. B and T cells require clonal expression and specificity in order to respond specifically to antigen. Target cells, on the other hand, must be able to present as wide a range of antigen to the effector cells as possible. Degenerate peptide binding, along with the expression of multiple, highly allelic forms, as with the MHC, accomplishes this.

The mechanisms used for the diversification of these antigen-recognition gene systems operate at the DNA, RNA, and protein levels. This diversity is represented in both the individual and the population as a vast array

of different types of molecules employing very different antigen-binding strategies. Most importantly, this ability to respond so specifically to the universe of antigens indicates the tremendous flexibility of the immunoglobulin homology unit in defining both recognition and functional operations.

Features of the Immunoglobulin Gene Superfamily

The immunoglobulin gene superfamily now has more than 40 different identified genes or gene families that fall into a variety of different evolutionary and functional categories (Fig. 1) (Williams 1987; Hunkapiller and Hood 1989). They are expressed by a variety of cell types, including lymphocytes (with an especially large number on T cells), nerve cells, various nervous-system-associated cells, bone-marrow-derived cells, fibroblasts, etc. The superfamily receptors fall into three general categories defined by how they associate with their ligand: homotypic (interaction with an identical subunit), heterotypic (interaction with a nonidentical homology unit), and a ligand without a homology unit. The recognition function of these receptors may be linked to various effector functions such as cellular differentiation or the activation of previously programmed cellular functions. In other cases, the recognition function serves merely to fix cells in particular positions with respect to other cells through cross-linking of receptor and ligand.

The non-immune receptor members of the immunoglobulin gene superfamily, most of which are involved in molecular recognition at the cell surface, fall into several functionally definable categories (Fig. 5). Morphogenic factors facilitate surface interactions between cells that lead to changes in gene expression and changes in cell movement, shape, and function (e.g., N-CAM, fasciclin II, contactin, MAG, P_0, I-CAM). A second category is represented by kinase growth factor receptors, a specialized group associated with a tyrosine kinase domain (e.g., PDGF receptor, CSF-1 receptor). These molecules play a key role in the proliferation and/or differentiation of specific cell lineages by linking a cascade of intracellular reactions with the recognition of a specific triggering ligand. Third, the specialized differentiation factors include many unrelated molecules responsible for diversity of functions including cellular activation, differentiation, and migration. Many of these molecules facilitate even more complex interactions between other members of the superfamily (e.g., CD3, CD4, CD8). Fourth, the antibody-binding/transport proteins attach to particular C-

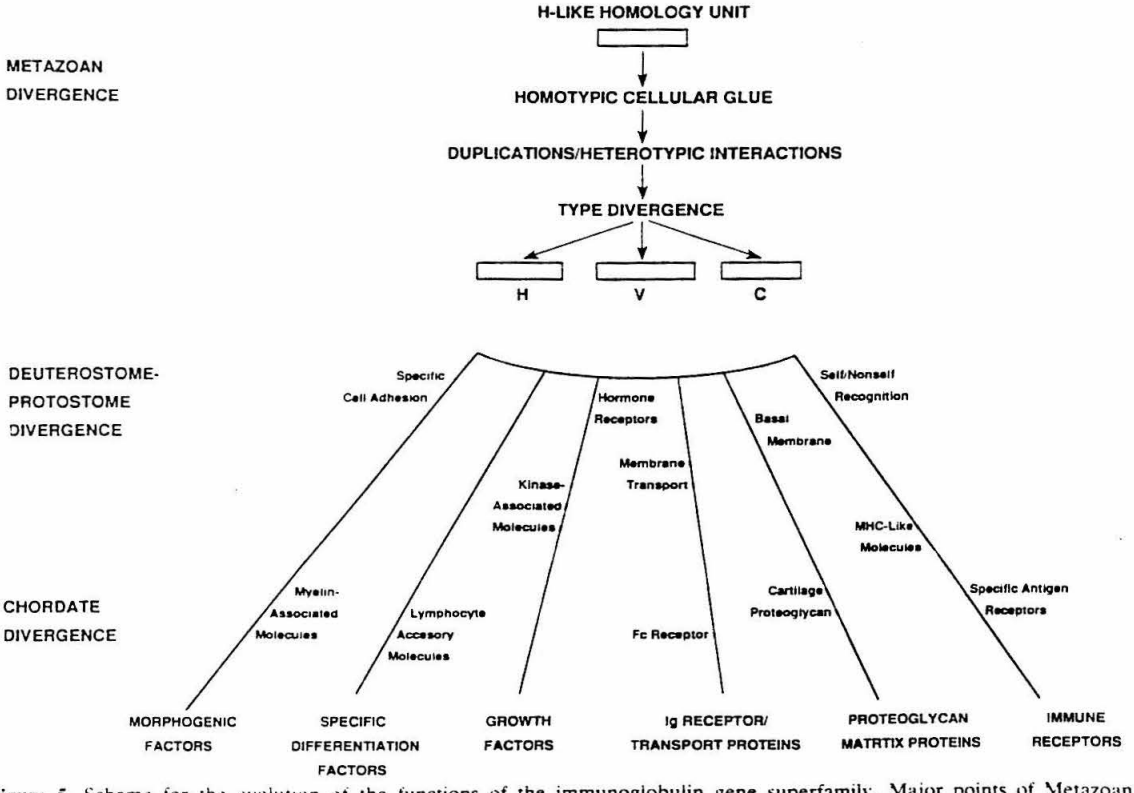


Figure 5. Scheme for the evolution of the functions of the immunoglobulin gene superfamily. Major points of Metazoan divergences are given at the left. Smaller print functions are associated with one of the five major lines of functional categories now evident in the immunoglobulin gene superfamily (see text). The arc is to emphasize that this is not meant to be considered a genealogic tree.

region domains in order to immobilize humoral antibody or ferry specific isotypes across particular membrane/cellular boundaries (e.g., Fc receptors, poly[Ig] receptor). Finally, the proteoglycan-associated proteins stabilize complex connective matrices such as cartilage (e.g., link protein, heparin sulfate proteoglycan [HSP]). They must cross-link other molecules, hence they have multiple binding moieties. The functions of many immunoglobulin superfamily molecules are not known (e.g. CD1, Thy-1, MRC OX2, CEA, Qa, Tla, CD7, CD28, B29). However, most would appear to fall into one of these categories.

Analysis of the non-immune receptor members of the immunoglobulin superfamily reveals that there is a predominance of polyunit structures (Fig.1); that is, most of the molecules are constructed from a series of tandemly repeated homology units. Comparisons at the sequence level indicate that most of these polyunit structures have arisen independently; they are not all the descendants of a primordial polyunit gene (Hunkapiller and Hood 1989). That the same organizational motif arose so many times means both that it must have significant advantages in many different contexts and that it spontaneously appears often enough for selective processes to operate. The general encoding of homology units in discrete exons and the conservation of symmetric splicing rules means that single genetic events can expand, contract, or shuffle any number of exons and still generate an in-frame transcript, mitigating to some degree possible negative selective issues. Therefore, with simple recombinational events, valency for ligand can be multiplied, and novel functional and recognition associations can be generated (Fig. 6).

Essentially all of the polyunit sequences are made up primarily of H units. With the exception of immunoglobulin heavy chains, none of the known homology unit polypeptides have more than a single V or C unit. However, both can be in array with units of a different type (Fig. 1). Also, V and C homology units usually are found in dimeric configurations, generating quaternary domains, whereas there are no known examples of dimeric H-unit proteins. These differences between H proteins and those with V and C units may reflect a greater ability of H units to pack together into a polydomain structure or, alternatively, not pack at all—to exist as open half-domain structures. This latter configuration is consistent with a model of polyunit receptors that interact with their ligands through domain-like interactions of complementary homology units. In support of this idea, essentially all characterized receptor-ligand operations between members of the immunoglobulin superfamily involve at least one, and usually two, H-unit proteins. In either case, it seems that the H unit retains the most flexibility for interaction with different homology units, probably representing the earliest, most basic form of the homology motif.

Two particular examples of the morphogenic members demonstrate the antiquity of the homology unit. Fasciclin II and amalgam are cell adhesion molecules involved in the development of the nervous systems of

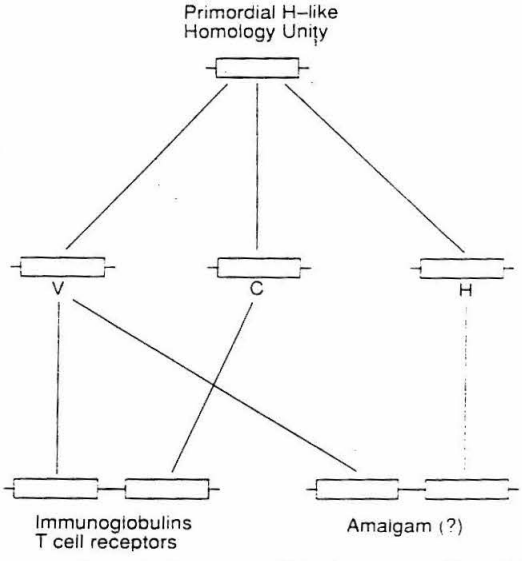


Figure 6. Exon shuffling possibilities between V, H, and C homology units in the immunoglobulin gene superfamily (see Fig. 1 and text).

grasshoppers and *Drosophila*, respectively, in a manner at least analogous to the vertebrate nervous system adhesins, N-CAM and myelin-associated glycoprotein (MAG) (Harrelson and Goodman 1988; Seeger et al. 1988). Both are polydomain sequences with multiple H homology units (Fig. 1), indicating an ancient association of the H motif with morphogenesis. Amalgam may also include a single V motif, illustrating both the early specialization of the V motif and the occurrence of exon shuffling. This means that the homology unit predated the divergence of protostomes and deuterostomes that occurred over 600 million years ago.

Another characteristic of the immunoglobulin superfamily is the association of homology unit regions with other, unrelated functional motifs, presumably through exon shuffling events. Among the best-characterized representatives are the kinase growth factor receptors for PDGF and CSF-1 (Lai et al. 1987). Each has an extracellular array of H homology units for receptor-ligand interaction and an intercellular domain that belongs to the tyrosine kinase superfamily. Also, fibronectin-binding domains are found in various morphogenic members. These examples share a recurring organizational theme, a specific receptor region composed of homology units, and an unrelated reactive moiety. These exon shuffling events broaden the biological contexts in which the homology unit functions.

Evolution of the Immunoglobulin Gene Superfamily

We believe that the immunoglobulin gene superfamily emerged with the radiative evolution of Metazoan organisms when the requirements for increased cell-

surface molecular recognition were expanding rapidly, presumably in association with increasing tissue specialization. With this emergence, the gene encoding the primordial homology unit was duplicated many times in many different contexts. This resulted in an explosion of different molecules involved in numerous forms of cell-surface recognition that paralleled the development of Metazoan complexity. Perhaps the most primitive of these functions is evident in modern morphogenic factors.

It is important to emphasize that molecules grouped together by function as in Figure 5 are not necessarily directly linked evolutionarily. In fact, over time, molecules with evolutionary roots in one group can acquire the attributes of another functional system. The development of the V, C, and H homology units seems to have coincided with this radiation of functional groups and the development of Metazoan complexity. The extreme divergence between homology unit sequences belonging to different groups as well as to the same group testifies to the early divergence of the distinct motifs. Indeed, the enormous number of homology unit representatives resulting from this expansion and their current distant relatedness makes the determination of their exact lineage relationships virtually impossible except in the broadest contexts. The genealogical tree for molecules containing homology units is further complicated by the ability of different, unrelated homology unit proteins to interact with each other in novel combinatorial associations (e.g., MHC class I chain and β_2-microglobulin) and the exon shuffling allowed by conserved splicing rules (Fig. 6). For example, it is not likely that V and C units have always been tandemly linked since their divergence from the precursor unit, or they would most likely be more similar to each other than either is to the H unit. MRC OX2, which has a VC structural organization similar to that of T-cell receptor chains, may be a paradigm of the precursor molecule generated by the exon linkage of separate V and C homology units before the evolution of rearrangement (Johnson and Williams 1986). Hence, any detailed evolutionary tree drawn for the members of the immunoglobulin gene superfamily must be considered significantly speculative.

It is important to view the evolution of the immunoglobulin gene superfamily as the result of preadaptive events that generated molecules from precursors that often had no functional relatedness to the descendant molecules. We must be careful, consequently, in developing ad hoc explanations for the evolution of phenotypic systems that involve complex associations between many members of the immunoglobulin superfamily, such as the systems of antigen recognition. It is highly unlikely that the molecules involved in such systems coevolved after their emergence within the context of only one system. They more likely were recruited from many different precursor origins, converging toward an ever more complex association, rather than diverging from simple origins to fill the "needs" of an emerging phenotype. Such coalescence of immunoglobulin superfamily members was probably facilitated by the tendency of different homology units to interact, establishing the associative possibilities from which complex interactions could arise. Given these caveats, however, we believe that simple generalizations can still be made as to the general flow of the evolution of the immunoglobulin gene superfamily (Fig. 5).

The existence of homology units in invertebrates such as insects is consistent with the idea that the immunoglobulin gene superfamily emerged early in Metazoan evolution. It is likely that the original homology unit exhibited homotypic interactions with homology units on other like cells, perhaps acting as nothing more than a cellular glue (Williams 1982) in a simple, nondifferentiated, multicellular aggregate. The homophyllic character of morphogenic molecules such as the neural cell adhesin, N-CAM, supports the feasibility of this scenario. The duplication of this early homology unit gene enabled the accumulation of a wide variety of novel mutations within the basic structure of the homology unit. This variation generated markers which, with the development of changes in the timing of expression, evolved to become tissue-specific, supporting ever more complex interactions in morphogenesis. Alternatively, of course, it could have been that cellular specialization arose prior to true multicellularity and that it was the communication needs between such cells that drove the emergence of the Metazoa and ultimately the diversity of the homology unit proteins. In either case, this required that early in Metazoan evolution, molecular recognition by homology domains became tied to activation signals that transmitted information from outside to inside the cell. For example, the association of homology units with a tyrosine kinase domain was probably the result of exon shuffling events, linking a specific receptor/ligand interaction to established activation pathways.

The ability of the homology unit to tolerate a wide range of variability no doubt allowed significant allelic variation to arise and generate in effect genotypic or self-markers (and conversely, non-self-markers). A test for non-self could have initially proved useful in favoring the fusion of heterogenetic gametes during sexual reproduction (Burnet 1971; Monroy and Rosati 1979). Members of the immunoglobulin superfamily have been implicated in this process in mammals (Lyon 1984; V. Scofield, pers. comm.). However, the ability to recognize and respond negatively to non-self is also important to any Metazoan organism that competes with its own kind at some time in its life cycle for space or substrate, particularly those that expand through asexual budding. This would require that heterotypic homology unit interactions representing allogeneic interactions become linked to potentially destructive processes. Interestingly, tunicates have a highly polymorphic gene system apparently responsible for both gametic exclusion and histocompatibility (Scofield et al. 1982). Histoincompatibility reactions have been recorded in most multicellular animal phyla (Hildemann

et al. 1980), raising the possibility that MHC-like genes were integral to the development of complex Metazoa.

Histocompatibility mechanisms that respond to foreign tissues could presumably also respond to any altered self-markers, for example, arising from the interactions of self-histocompatibility molecules with other macromolecular structures. This suggests that the specific immune response is an outgrowth of these histocompatibility defenses. Early in the evolution of self-defense mechanisms, presumably existing phagocytic operations became linked and targeted to the allogeneic response by the association of the recognition elements with the activation of phagocytic processes, perhaps in a manner analogous to the exon shuffling seen in growth factor receptors. These processes may have represented the first forms of antigen processing. It is possible that MHC-like molecules or at least the precursor to the peptide-binding domain already served as some kind of membrane transport protein that became linked to this process. Facilitated by their membrane cycling, these MHC-like molecules eventually become the receptors for foreign peptide fragments. The ability to present allogeneic antigen on host cells tied to secondary signaling processes allowed an amplification of the antigenic signal. Transport of that signal away from the site of allogeneic contact could recruit by activation even more effector cells. This could explain the present linkage of the MHC and T-cell receptor antigen recognition systems. The acquired ability of MHC molecules to bind and present peptides as "altered self" leads directly to the problem of how the primitive allogeneic recognition system could distinguish MHC complexes with self versus non-self peptides. Although the solution to the problem is not clear, it must involve additional signaling between the antigen-presenting cell and the effector cell. It is also possible that early antigen presentation was restricted to only specific proteins, limiting the problem of tolerance substantially. The example of the highly specific binding and transport of maternal milk antibodies by an Fc receptor that is directly homologous to class I sequences may serve as a paradigm in this regard (Simister and Mostov 1989).

It is probable that the cell-cell interactions involved in morphogenesis arose before the development of immune-type reactions. This branching was no doubt a major event in the evolution of the family, each branch taking advantage of the ability of the homology unit to accommodate variability in a fundamentally different way. Members of the morphogenic differentiation branches became more and more specialized and targeted in their interactions, generating enormous varieties of highly specialized molecules. The immune branch, on the other hand, became more and more dependent on the variation possibilities within one family of genes, allowing that family to have the broadest possibilities for interaction while retaining the highly specific interactions of the individual receptors.

A key event in the evolution of antigen recognition receptors was the acquisition of the ability to rearrange gene segments. A likely scenario is that a transposon inserted into a primordial V gene separating the initial V and J gene segments (Sakano et al. 1979). Transposon excision with the subsequent joining of the separated elements eventually came under developmental control (e.g., a limited window of time within which the enzymes responsible for rearrangement are active). However, this later process was never precise, generating more nonproductive than productive rearrangements. This had the effect of "excluding" any nonproductively rearranged V gene from generating a functional protein. With only a limited number of V to J possibilities, this would be functionally the equivalent of allelic exclusion because the likelihood of more than one productive rearrangement in one cell would be small. Thus, a single cell could produce only one or a limited number of antigen-specific receptors. With near-clonal expression, the process of clonal selection by foreign macromolecular patterns could now occur, generating selective pressure for ever finer control of allelic expression. The lower the rate of productive rearrangements per number of V-J rearrangement possibilities, the greater the number of such potential pairings allowable while still maintaining near-clonal expression. Therefore, as the control mechanisms for allelic exclusion became more defined, the number of gene segments could be expanded, thereby generating even more receptor diversity potential.

A T-cell-like recognition system probably evolved before that of B cells. It maintained its earlier linkage to non-self cell destruction and thus the recognition of antigen in the context of MHC molecules. Duplications of gene families occurred to generate the various types of T-cell receptors and ultimately the B-cell receptor gene families. Central to this would have been freeing B cells from direct cell-cell interaction requirements and thus the need for MHC binding.

Examination of the antibody genes and diversity in the shark *Heterodontus*, the chicken, and mammals illustrates three different stages in the evolution of rearranging multigene families (Fig. 7). The shark has multiple clusters of tandemly repeated V-D-J-C genes for heavy chains (Kokubu et al. 1987). It is unclear whether V gene segments can only rearrange to their tandemly associated J gene segment, but such possibilities at least seem limited, reducing the combinatorial possibilities. However, junctional variation is still possible. Increase in the effective total number of V-J-C sets possible in one genome is probably limited as well

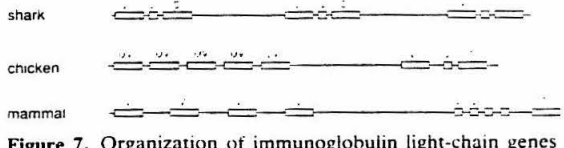


Figure 7. Organization of immunoglobulin light-chain genes in the shark, chicken, and mammal (see Fig. 4 and text).

by the consequent increased likelihood of multiple productive rearrangements, assuming that allelic exclusion has some stochastic quality. These constraints are consistent with the fact that cartilaginous fish have a more reduced expressed antibody diversity than other vertebrates (Litman et al. 1982). The chicken, on the other hand, has a complex antibody repertoire, but has only a single V and J gene segment for both light and heavy chains, eliminating all combinatorials. However, the single V gene can be corrected against 25 or more closely linked pseudo-V_L gene segments by a process of random gene conversion, thus generating enormous diversity (Weill 1986). Certainly, allelic exclusion is much easier in such a system. Allelic variation may play an important role in this system as a way to expand the range of diversity (presumably, the more divergent are two homologs, the greater the difference in the possible antigens that can be surveyed by one animal). The most sophisticated potential for diversity is seen in the mammals. The grouping of V, D, and J gene segments should allow even tighter control on allelic exclusion, because each nonproductive rearrangement actually reduces the remaining possible rearrangement target sites. The number of possible V-J pairings is limited by the element with the smallest number of copies, usually D or J. The remaining element(s) can expand without impact on allelic exclusion. Hence, with this organization there is much less restriction on the expansion of the number of V gene segments and the consequent increase in combinatorial and junctional variation potential. (It is interesting to note that allelic exclusion of T-cell receptor α chains may not be as complete as other chains [Malissen et al. 1988]. This may correlate with the fact that the number of J gene segments far outnumbers that of any other chain, approaching a substantial fraction of the size of the V_α gene segment number. However, experiments with transgenic animals suggest that specific, nonstochastic regulatory mechanisms also play a role in allelic exclusion [Blüthmann et al. 1988].) As diversity and its range increase in a gene segment family, the importance of allelic variation is probably reduced.

With the expansion of V gene segments, however, it becomes more important to link their transcription to rearrangement. In the case of the mouse V_H family, for example, the productively rearranged V_H gene is derived from one of over a thousand (including both chromosomes) V_H gene segments (Livant et al. 1986), each with its own transcription initiation site. However, effective transcription is possible only when rearrangement brings a V gene segment into the vicinity of specific enhancer sequences near the C gene. Thus, the evolution of rearrangement probably had more to do with regulation of allelic exclusion and expression than diversity per se (Hunkapiller and Hood 1989). However, the development of clonal expression is what freed the constraints on V gene expansion and facilitated the evolution of diverse antigen-specific repertoires.

Implications of the Immunoglobulin Gene Superfamily

The history of the immunoglobulin gene superfamily reflects both the fundamental diversifying and combinatorial properties of the immunoglobulin homology unit. New functional possibilities arise by alteration of the informational units involved (e.g., nucleotides, exons, genes, and entire multigene families), by shuffling of these elements, and by the linkage of homology units with other, independent functional processes. The duplication of a multigene family and the attendant *cis*-acting control mechanisms can in a single event create the genetic basis for a complex new phenotype, thereby increasing the possibility for very rapid evolutionary change. The evolution of the rearranging gene families illustrates this principle clearly. On the assumption that the development of rearrangement occurred only once, the proto-immune receptor was probably a homodimer (allowing for allelic forms). Duplication and subsequent divergence of the proto-family led then to the modern heterodimeric form and, consequently, a whole new diversifying principle. These duplications promoted the creation of genes whose products could interact with one another, generating the combinatorial potentials between the products of two gene families. Multiple pairs of specifically interacting gene families would coevolve over time. Changes in the regulation of these pairs would eventually generate the distinctive T- and B-cell receptors seen today. It seems probable that the development or addition of antibody effector domains arose after the regulatory separation of the B- and T-cell lineages. Thus, the duplication of multigene families can generate new multigene families with entirely distinct functional potential. The duplication of entire or parts of multigene families, therefore, provides the potential for extremely rapid evolutionary change and the relatively saltational acquisition of complex new phenotypes. This is consistent with recent models of evolution that propose that more punctuated events may be stronger drivers of large-scale phenotypic change than the gradualism of simple point mutations of structural genes (see Gould 1989).

The expansion of the immunoglobulin gene superfamily to carry out a multitude of different major functions in eukaryotic biology is a classic example of evolutionary canalization. This is the process whereby once an evolutionary branching has been made, a de facto direction for future evolutionary events has been made more likely, because the path taken at one juncture limits the future paths available (i.e., it is hard to go backwards). The early success of the homology unit signaled an important early juncture of Metazoan development as regards the possible mechanisms and motifs employed in cell-cell interactions. Also, the tendency of homology units to interact drives the development of ever more complex associations of proteins containing homology units. This complexity sub-

Figure 8. A single homeotic mutation creates a four-winged fly from the wild-type two-winged fly (courtesy of E. Lewis).

sequently generates the building blocks of additional, complex phenotypes. In this regard, the immunoglobulin homology unit motif is repeated in many different variations and contexts, tying together in an evolutionary and sometimes functional sense many areas of eukaryotic biology. Particularly fascinating are the members of the superfamily that are expressed by both the immune and nervous systems (see Parnes and Hunkapiller 1987). This coexpression raises the possibility of shared receptor molecules that facilitate communication between these two complex systems that deal with external stimuli. Indeed, it seems likely that the nervous system plays an important role in regulating the immune response, perhaps mediated by interactions between immunoglobulin superfamily members.

The evolution of control mechanisms is one of the real keys to understanding how the complex phenotypes of eukaryotes have arisen. However, a full understanding of the evolution of regulatory mechanisms is impossible until we understand more of the functional details of current regulatory mechanisms. Hence, the study of developmental biology and evolutionary biology are two sides of the same coin. How is it that a new multigene family can be created and in time its members be expressed in different cell types within different functional contexts? In this regard, it is interesting to consider the evolutionary processes that led to the development of the modern fly from the early annelids. The more specialized descendants of a simple segmental creature had to acquire the developmental programs that led to ever-increasing specialization of their segments (e.g., head, thorax, and abdomen) and their appendages, such as legs and wings. Presumably, these features are encoded by coordinately expressed linked and unlinked genes as well as gene families that must be duplicated and controlled. It seems likely that the required developmental changes were more the result of regulation and timing changes than major modifications to the structural genes. Indeed, a single homeotic mutation can convert a modern two-wing fly to a more ancient four-wing form (Fig. 8), indicating that the genetic information for generating older phenotypes is still contained intrinsically within contemporary genomes and that it is the context of its expression that determines the final phenotype, not the coding information itself. Such changes in homeotic and multigene systems illustrate that changes in the regulation of complex traits may be more likely to generate selectively tolerable phenotypic changes than direct major alteration of structural genes, thus promoting relatively saltational changes in phenotype.

The immunoglobulin gene superfamily has given us a glimpse into the awesome complexities of modern biology; at the same time, its study offers hope for unraveling some of these complexities. Questions concerning the overall diversity of the immunoglobulin gene superfamily and whether other complex eukaryotic systems employ gene superfamilies of similar complexity await future investigations.

ACKNOWLEDGMENTS

The National Institutes of Health has supported this work. We thank Sue Lewis and all those of the typing staff for help with this paper. B.F.K. is the recipient of a fellowship from the Sloan Foundation.

REFERENCES

Amzel, L.M. and R. J. Poljak. 1979. Three-dimensional structure of immunoglobulins. *Annu. Rev. Biochem.* **48:** 961.

Bank, I., R.A. DePinho, M.B. Brenner, J. Caasimeris, F.W. Alt, and L. Chess. 1986. A functional T3 molecule associated with a novel heterodimer on the surface of immature human thymocytes. *Nature* **322:** 179.

Barth, R.K., B.S. Kim, N.C. Lan, T. Hunkapiller, N. Sobiek, A. Winoto, H. Gershenfeld, C. Okada, D. Hansburg, I. Weissman, and L. Hood. 1985. The murine T-cell receptor employs a limited repertoire of expressed V_β gene segments. *Nature* **316:** 517.

Blüthmann, H., P. Kisielow, Y. Uematsu, M. Malissen, P. Krimpenfort, A. Berns, H. von Boehmer, and M. Steinmetz. 1988. T-cell specific deletion of T-cell receptor trans-

genes allows functional rearrangement of endogenous alpha and beta genes. *Nature* **334:** 156.

Brenner, M.B., J. McLean, D.P. Dialynas, J.L. Strominger, J.A. Smith, F.L. Owen, J.G. Seidman, S. Ip, F. Rosen, and M.S. Krangel. 1986. Identification of a putative second T-cell receptor. *Nature* **322:** 145.

Burnet, F.M. 1971. "Self-recognition" in colonial marine forms and flowering plants in relation to the evolution of immunity. *Nature* **232:** 230.

Capra, K. and J.M. Kehoe. 1974. Variable region sequences of 5 human immunoglobulin heavy chains of the variable heavy chain. III. Subgroup definitive identification of 4 heavy chain hypervariable regions. *Proc. Natl. Acad. Sci.* **71:** 845.

Chien, Y.-H., M. Iwashima, K.B. Kaplan, J.F. Elliot, and M.M. Davis. 1987. A new T-cell receptor gene located within the α locus and expressed early in T-cell differentiation. *Nature* **327:** 677.

Clarke, S.H., K. Huppi, D. Ruezinsky, L. Standt, W. Gerhard, and W. Weigert. 1985. Inter- and intraclonal diversity in the antibody response to influenza hemagglutinin. *J. Exp. Med.* **161:** 687.

Claverie, J.-M. and P. Kourilsky. 1987. The peptidic self model: A reassessment of the role of the major histocompatibility complex molecules in the restriction of the T-cell response. *Ann. Immunol.* **137:** 425.

Cunningham, B.A., J.J. Hemperly, B.A. Murray, E.A. Prediger, R. Brackenbury, and G.M. Edelman. 1973. Neural cell adhesion molecule structure, immunoglobulin-like domains, cell surface modulation and alternative RNA splicing. *Science* **236:** 799.

Davis, M.M. 1985. Molecular genetics of the T-cell receptor β chain. *Annu. Rev. Immunol.* **3:** 537.

Davis, M.M., S.K. Kim, and L. Hood. 1980. Immunoglobulin class switching developmentally regulated DNA rearrangements during differentiation. *Cell* **22:** 1.

Early, P., J. Rogers, M. Davis, K. Calame, M. Bond, R. Wall, and L. Hood. 1980. Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways. *Cell* **30:** 313.

Edmundson, A.B., K.R. Ely, E.E. Abola, M. Schiffer, and N. Panagiotopoulos. 1975. Rotational allomerism and divergent evolution of domains in immunoglobulin light chains. *Biochemistry* **14:** 3953.

Eisen, H.N. 1986. Why affinity progression of antibodies during immune responses is probably not accompanied by parallel changes in the immunoglobulin-like antigen-specific receptors on T cells. *BioEssays* **4:** 269.

Elliot, J.F., E.P. Rock, P.A. Patten, M.M. Davis, and Y.-H. Chien. 1988. The adult T-cell receptor δ chain is diverse and distinct from that of fetal thymocytes. *Nature* **331:** 627.

Gearhart, P.M. and D.F. Bogenhagen. 1983. Clusters of point mutations are found exclusively around rearranged antibody variable genes. *Proc. Natl. Acad. Sci.* **80:** 3437.

Gearhart, P.M., N.D. Johnson, R. Douglass, and L. Hood. 1981. IgG antibodies to phosphorylcholine exhibit more diversity than their IgM counterparts. *Nature* **291:** 29.

Germain, R.N. and B. Malissen. 1986. Analysis of the expression and function of class II major histocompatibility complex-encoded molecules by DNA-mediated gene transfer. *Annu. Rev. Immunol.* **4:** 281.

Gould, S.J. 1989. *Wonderful life: The Burgess Shale and the nature of history*. W.W. Norton, New York.

Goverman, J., T. Hunkapiller, and L. Hood. 1986. A speculative view of the multicomponent nature of T-cell antigen recognition. *Cell* **45:** 475.

Hannum, C., J.H. Freed, G. Tarr, J. Kappler, and P. Marrack. 1984. Biochemistry and distribution of the T-cell receptor. *Immunol. Rev.* **81:** 161.

Harrelson, A.L. and C.S. Goodman. 1988. Growth cone guidance in insects: Fasciclin II is a member of the immunoglobulin supertamily. *Science* **242:** 700.

Hayday, A.C., H. Saito, S.D. Gillies, D.M. Kranz, G. Tanigawa, H.N. Eisen, and S. Tonegawa. 1985. The structure organization and somatic rearrangement of T-cell γ genes. *Cell* **40:** 259.

Hildemann, W.H., C.H. Bigger, J.S. Johnston, and P.L. Jokiel. 1980. Characteristics of immune memory in invertebrates. *Dev. Immunol.* **10:** 9.

Hill, R.L., R. Delaney, R.E. Fellow, and H.E. Lebowitz. 1966. The evolutionary origins of the immunoglobulins. *Proc. Natl. Acad. Sci.* **56:** 1762.

Honjo, T. 1983. Immunoglobulin genes. *Annu. Rev. Immunol.* **1:** 499.

Honjo, T. and S. Habu. 1985. Origin of immune diversity genetic variation and selection. *Annu. Rev. Biochem.* **54:** 803.

Hood, L., M. Kronenberg, and T. Hunkapiller. 1985. T-cell antigen receptors and the immunoglobulin supergene family. *Cell* **40:** 225.

Hood, L., M. Steinmetz, and B. Malissen. 1983. Genes of the major histocompatibility complex of the mouse. *Annu. Rev. Immunol.* **1:** 529.

Hunkapiller, T. and L. Hood. 1989. Diversity of the immunoglobulin gene superfamily. *Adv. Immunol.* **44:** 1.

Ioannides, C.G., K. Itoh, F.E. Fox, R. Pahwa, R.A. Good, and C.D. Platsoucas. 1987. Identification of a second T-cell antigen receptor in human and mouse by an antipeptide γ-chain-specific monoclonal antibody. *Proc. Natl. Acad. Sci.* **82:** 7701.

Johnson, P. and A.F. Williams. 1986. Striking similarities between antigen receptor J pieces and sequence in the second chain of the murine CD8 antigen. *Nature* **323:** 74.

Kim, S., M. Davis, E. Sinn, P. Patten, and L. Hood. 1981. Antibody diversity: Somatic hypermutation of rearranged V_H genes. *Cell* **27:** 573.

Klein, J. 1982. *Immunology: The science of self-nonself discrimination*. Wiley, New York.

Klein, J. and F. Figueroa. 1981. Polymorphism of the mouse H-2 loci. *Immunol. Rev.* **60:** 23.

———. 1986. The evolution of class I MHC genes. *Immunol. Today* **7:** 41.

Kokubu, F., K. Hinds, R. Litman, M.S. Shamblott, and G.W. Litman. 1987. Extensive families of constant region genes in a phylogenetically primitive vertebrate indicate an additional level of immunoglobulin complexity. *Proc. Natl. Acad. Sci.* **84:** 5868.

Kronenberg, M., G. Siu, L. Hood, and N. Shastri. 1986. The molecular genetics of the T-cell antigen receptor and T-cell antigen recognition. *Annu. Rev. Immunol.* **4:** 529.

Lai, C., M.A. Brow, K.-A. Nave, A.B. Noronha, R.H. Quarles, F.E. Bloom, R.J. Milner, and J.G. Sutcliffe. 1987. Two forms of 1B236/myelin-associated glycoprotein, a cell adhesion molecule for post-natal neural development are produced by alternative. *Proc. Natl. Acad. Sci.* **84:** 4337.

Litman, G.W., J. Stolen, H.O. Sarvas, and O. Mäkelä. 1982. The range and fine specificity of the anti-hapten immune response: Phylogenetic studies. *J. Immunogenet.* **9:** 465.

Livant, D., C. Blatt, and L. Hood. 1986. One heavy chain variable region gene segment subfamily in the BALB.c mouse contains 500–1000 or more members. *Cell* **47:** 461.

Luzzati, A.L., I. Lefkovits, and B. Pernis. 1973. Homogeneity of antibodies produced by clones *in vitro*. *Eur. J. Immunol.* **3:** 636.

Lyon, M.F. 1984. Transmission radio distortion in mouse t-haplotypes is due to multiple distorter genes acting on a responder locus. *Cell* **37:** 621.

Malissen, M., J. Trucy, F. Letourneur, N. Rebai, D.E. Dunn, F.W. Fitch, L. Hood, and B. Malissen. 1988. A T-cell clone expresses two T-cell receptor alpha genes but uses one alpha-beta heterodimer for allorecognition and self MHC-restricted antigen recognition. *Cell* **55:** 49.

Matis, L.A., R. Cron, and J.A. Bluestone. 1987. Major histocompatibility complex-linked specificity of $\gamma\delta$ receptor-bearing T lymphocytes. *Nature* **330:** 262.

Mengle-Gaw, L. and H.O. McDevitt. 1985. Genetics and expression of Ia antigens. *Annu. Rev. Immunol.* **3:** 367.

Meuer, S.C., O. Acuto, T. Hercend, S.F. Schlossman, and E.L. Reinherz. 1984. The human T-cell receptor. *Annu. Rev. Immunol.* **2:** 23.

Monroy, A. and F. Rosati. 1979. The evolution of the cell-cell recognition system. *Nature* **278:** 165.

Novotny, J., S. Tonegawa, H. Saito, D. Kranz, and H. Eisen. 1986. Secondary, tertiary, and quaternary structure of T-cell specific immunoglobulin-like polypeptide chains. *Proc. Natl. Acad. Sci.* **83:** 742.

Parnes, J.R. and T. Hunkapiller. 1987. L3T4 and the immunoglobulin gene superfamily: New relationships between the immune system and the nervous system. *Immunol. Rev.* **100:** 109.

Sablitzky, F., G. Wildner, and K. Rajewsky. 1985. Somatic mutation and clonal expansion of B cells in an antigen-driven immune response. *EMBO J.* **4:** 345.

Sakano, H., K. Huppi, G. Heinrich, and S. Tonegawa. 1979. Sequences of the somatic recombination sites of immunoglobulin light chain genes. *Nature* **280:** 288.

Schwartz, R.H. 1984. The role of gene products of the major histocompatibility complex in T-cell activation and cellular interactions. In *Fundamental immunology* (ed. W.E. Paul), p. 379. Raven Press, New York.

Scofield, V.L., J.M. Schlumpberger, L.A. West, and I.L. Weissman. 1982. Protochordate allorecognition is controlled by a MHC-like gene system. *Nature* **295:** 499.

Seeger, M.A., L. Haffley, and T.L. Kaufman. 1988. Characterization of amalgam, a member of the immunoglobulin superfamily from *Drosophila*. *Cell* **55:** 589.

Simister, N.E. and K.E. Mostov. 1989. An Fc receptor structurally related to MHC class I antigens. *Nature* **337:** 184.

Sprent, J. and S.R. Webb. 1987. Function and specificity of T cell subsets in the mouse. *Adv. Immunol.* **41:** 39.

Strominger, J.L., H.T. Orr, P. Parham, H.L. Ploegh, D.L. Mann, H. Bilofsky, H.A. Saroff, T.T. Wu, and E.A. Kabat. 1980. An evaluation of the significance of amino-acid sequence homologies in human histocompatibility antigens HLA-A and HLA-B with immunoglobulins and other proteins using relatively short sequences. *Scand. J. Immunol.* **11:** 573.

Tonegawa, S. 1983. Somatic generation of antibody diversity. *Nature* **302:** 575.

Weill, J.C. 1986. Generation of diversity in the avian bursa. *Prog. Immunol.* **6:** 20.

Williams, A.F. 1982. Surface molecules and cell interactions. *J. Theor. Biol.* **98:** 221.

———. 1984. The immunoglobulin superfamily takes shape. *Nature* **308:** 12.

———. 1987. A year in the life of the immunoglobulin superfamily. *Immunol. Today* **8:** 298.

Wu, T.T. and E.A. Kabat. 1970. An analysis of the sequences of the variable regions of Bence-Jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med.* **132:** 211.

VI. FUNCTION AND STRUCTURE OF THE IMMUNE RECEPTORS

VI.1 T-cell Receptors

Cell, Vol. 37, 1101-1110, July 1984. Copyright © 1984 by MIT

0092-8674/84/071101-10 $02.00/0

Mouse T Cell Antigen Receptor: Structure and Organization of Constant and Joining Gene Segments Encoding the β Polypeptide

Marie Malissen,* Karyl Minard,* Shelley Mjolsness,* Mitchell Kronenberg,* Joan Goverman,* Tim Hunkapiller,* Michael B. Prystowsky,† Yasunobu Yoshikai,‡ Frank Fitch,† Tak W. Mak‡ and Leroy Hood*

*Division of Biology
California Institute of Technology
Pasadena, California 91125
†Department of Pathology
University of Chicago
Chicago, Illinois 60637
‡Department of Medical Biophysics
University of Toronto
Ontario Cancer Institute
Toronto, Canada M4X 1K9

Summary

The germ-line joining (J) gene segments and constant (C) genes encoding the β chain of the mouse T cell antigen receptor have been isolated on a single cosmid clone. There are two constant genes, $C_{\beta 1}$ and $C_{\beta 2}$, each associated with a cluster of J_β gene segments. The nucleotide sequences of the $C_{\beta 2}$ gene and of the $J_{\beta 2}$ cluster gene segments have been determined. The coding sequence of the $C_{\beta 2}$ gene is very similar to the sequence of a cDNA clone encoded by the $C_{\beta 1}$ gene. The $C_{\beta 2}$ gene has four exons; exon–intron structure does not obviously correspond to the functional domains of the protein. The $J_{\beta 2}$ gene segment cluster contains six functional J gene segments. We have isolated specific probes for the $C_{\beta 1}$, $C_{\beta 2}$, $J_{\beta 1}$, and $J_{\beta 2}$ regions to examine DNA rearrangements in T lymphocytes. DNA rearrangements can occur in both J_β gene segment clusters, and both C_β genes appear functional.

Introduction

The vertebrate immune response employs two categories of antigen receptors that are expressed on the surface of lymphocytes, immunoglobulins on B cells and the antigen-specific receptors on T cells. The extensive diversity of both receptor repertoires suggests that B cells and T cells may employ similar strategies for generating antigen-specific receptors. The immunoglobulins are composed of two distinct polypeptides, light and heavy, that fold into variable and constant domains. The variable domains recognize foreign antigen, and the constant domains are involved in a variety of effector functions. The variable region for light chains is encoded by two distinct gene segments (Bernard et al., 1978), variable (V_L) and joining (J_L), whereas the variable region for heavy chains is encoded by three different gene segments, V_H, diversity (D), and J_H (Early et al., 1980). These gene segments are joined by DNA rearrangement during B cell differentiation to generate complete V_L (V_L–J_L) and V_H (V_H–D–J_H) genes (Honjo, 1983; Tonegawa, 1983).

Until recently, the nature of the T cell antigen receptor was controversial (Kronenberg et al., 1983). The generation of antibodies specific for individual T cell lines (clonotypic reagents) from humans and mice has permitted the first biochemical characterization of the T cell antigen receptor (Acuto et al., 1983; Kappler et al., 1983; McIntyre and Allison, 1983; Samelson et al., 1983). The T cell receptor is composed of two disulfide bridged chains, α and β, ranging in molecular weight from 40,000 to 50,000. Peptide map analyses of these chains from several different receptors indicate that both α and β chains are composed of constant and variable peptides, presumably encoded by separate variable and constant region genes. Furthermore, there are virtually no peptides common to α and β chains. Thus they are likely to be encoded by two distinct gene families.

Several cDNA clones that encode the β chains of mouse and human T cell receptors have been isolated by subtractive and differential screening techniques (Hedrick et al., 1984a, 1984b; Yanagi et al., 1984). The T cell receptor cDNA clones exhibit sequence homology to the immunoglobulin V, J, and C regions, and the corresponding regions undergo DNA rearrangements during T cell differentiation (Hedrick et al., 1984b; Toyonaga et al., submitted). We have recently suggested that the β variable regions are encoded by three distinct gene segments—V_β, D_β, and J_β (Siu et al., 1984). In this paper we report the isolation of a mouse cosmid clone containing two C_β genes, denoted $C_{\beta 1}$ and $C_{\beta 2}$. Each C_β gene has a 5′-associated cluster of J gene segments. Therefore the genes encoding at least one chain of the T cell antigen receptor appear similar to the gene families encoding immunoglobulins. We have determined the complete nucleotide sequence of the $C_{\beta 2}$ gene and the $J_{\beta 2}$ gene segment cluster.

Results

Southern Blot Analyses of Mouse Liver DNA with the Mouse C_β cDNA Probe Suggest That There Are at Least Two C_β Genes

Southern blot analyses of mouse liver DNA (liver DNA is considered not rearranged for T cell receptor genes) hybridized with a C_β probe derived from the mouse T cell tumor RBL5 (Caccia et al., 1984) are given in Figure 1. Only two or three distinct bands are present upon analysis with various restriction enzymes. These data suggest that there are at least two C_β genes. Under low stringency hybridization conditions, no additional bands have been detected.

The C_β Genes Are Linked and Each Has an Associated Cluster of J Gene Segments

A cosmid library constructed with DNA from the livers of B10.WR7 inbred mice was screened with the constant

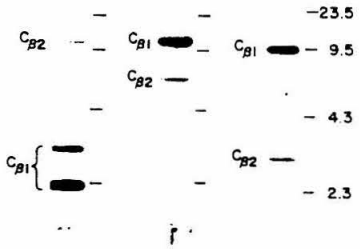


Figure 1. A Southern Blot Analysis of C57BL/6 Mouse Liver DNA with a $C_{\beta 1}$ Mouse cDNA Probe

The C57BL/6 DNA was digested with the restriction enzymes indicated. Bands known to contain the $C_{\beta 1}$ or $C_{\beta 2}$ genes from the intensity of hybridization and the analysis of the germ-line cosmid clone are indicated. The RBL5 cDNA probe used here contains most of the $C_{\beta 1}$ coding region and the 3′ untranslated sequence. As we discuss below, the coding regions of $C_{\beta 1}$ and $C_{\beta 2}$ are 96% similar. The difference in hybridization between $C_{\beta 1}$ and $C_{\beta 2}$ is the result of lack of similarity between the 3′ untranslated regions of $C_{\beta 1}$ and $C_{\beta 2}$. Size standards are given in kilobases for one blot; the positions of equivalent standards have been indicated for the others.

region portion of the human β chain cDNA YT35 (Yanagi et al., 1984). A hybridizing clone containing a 34 kb insert was isolated and characterized in detail. The restriction map of this DNA is shown in Figure 2. To locate potential coding sequences, Southern blots of the cloned genomic DNA were hybridized with a human constant region probe derived from YT35 or with synthetic oligonucleotides that correspond to portions of the YT35 cDNA sequence. The synthetic oligonucleotides (Horvath et al., submitted) include a 30-mer derived from the J sequence (nucleotides 412–441 of YT35), a 20-mer derived from the 3′ end of the constant region (nucleotides 757–776), a 17-mer derived from the connecting peptide (nucleotides 868–884), and an 18-mer derived from the transmembrane region (nucleotides 941–958). From the restriction map and Southern blots of the cosmid clone, we can draw the following conclusions. There are two constant region sequences separated by approximately 6 kb of DNA denoted $C_{\beta 1}$ and $C_{\beta 2}$ (Figure 2). Both C_β gene sequences are in the same 5′-to-3′ orientation. All of the hybridizing restriction fragments detected on Southern blots of liver DNA are also present on the cosmid (Figure 1). Therefore it appears that we have cloned all the C_β genes that are capable of readily cross-hybridizing with mouse or human C_β probes. Sequences located 5′ to both $C_{\beta 1}$ and $C_{\beta 2}$ genes hybridize with the oligonucleotide derived from the J region of the human cDNA clone. Partial nucleotide sequence analysis of a region 5′ to the $C_{\beta 1}$ gene indicates that multiple, functional J gene segments are present (data not shown).

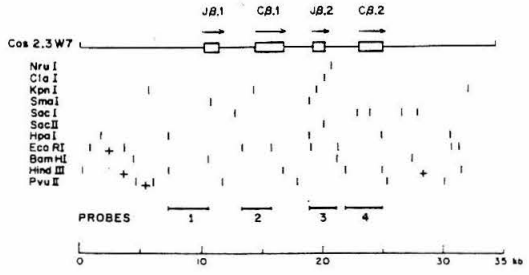


Figure 2. Restriction Map of the Mouse Cosmid Clone 2.3W7 Containing the $C_{\beta 1}$ and $C_{\beta 2}$ Genes with Their Associated J_β Gene Segments

Boxes indicate regions that hybridize to the human cDNA or to the synthetic oligonucleotides. The direction of transcription, indicated by arrows, was determined by hybridization with 5′ and 3′ probes for the C_β genes or by sequencing for the J_β segments. A (+) between two restriction sites indicates the presence of unassigned sites in this region for the same enzyme. Probes 1 through 4 are the following restriction fragments: (1) a 3.2 kb Hind III–Bam HI fragment; (2) a 2.5 kb Eco RI fragment; (3) a 2.3 kb Eco RI fragment, (4) a 3 kb Hind III fragment.

More complete nucleotide sequence of multiple functional $J_{\beta 2}$ sequences 5′ to the $C_{\beta 2}$ gene is presented below. In summary, the mouse genome contains two C_β genes, each of which is associated at its 5′ end with a cluster of J gene segments (Figure 2).

The $J_{\beta 2}$ Cluster Has Six Gene Segments

We have determined the nucleotide sequence of 2.3 kb of DNA beginning approximately 1.8 kb upstream of the $C_{\beta 2}$ gene and continuing in the 5′ direction. This region of the cosmid clone contains sequences that hybridize to the oligonucleotide corresponding to the J_β segment of the human cDNA sequence. Coding sequences in this region could be identified by several criteria. First, certain amino acid sequences are characteristic of almost all immunoglobulin and T cell receptor J gene segments. These nearly invariant amino acids include Phe-Gly-X-Gly-Thr, followed by a relatively hydrophobic sequence rich in leucine and valine (see Figure 5). Second, the nucleotide sequence that might encode a J gene segment could be checked for the requisite 5′ rearrangement signal and 3′ splice donor signal approximately 15–16 codons apart from one another. These signals are discussed further below. Using these criteria, six apparently functional $J_{\beta 2}$ gene segments have been identified (Figures 3A and 3B). These gene segments, spanning about 1100 nucleotides, are designated $J_{\beta 2.1}$ through $J_{\beta 2.6}$ from the 5′ end to the 3′ end of the cluster. The $J_{\beta 2}$ gene segments are separated by 40 to 300 nucleotides of intervening DNA sequence. We have sequenced about 400 nucleotides 5′ of the $J_{\beta 2.1}$ gene segment and 770 nucleotides 3′ of $J_{\beta 2.6}$ gene segment. Since the widest separation among the identified $J_{\beta 2}$ gene segments is 300 bases, it is likely we have identified all of the $J_{\beta 2}$ gene segments. Further sequence analysis in the 5′ and 3′ directions from the $J_{\beta 2}$ gene cluster is being carried out to verify this conclusion.

The $J_{\beta 2}$ gene segments range in size from 46 nucleo-

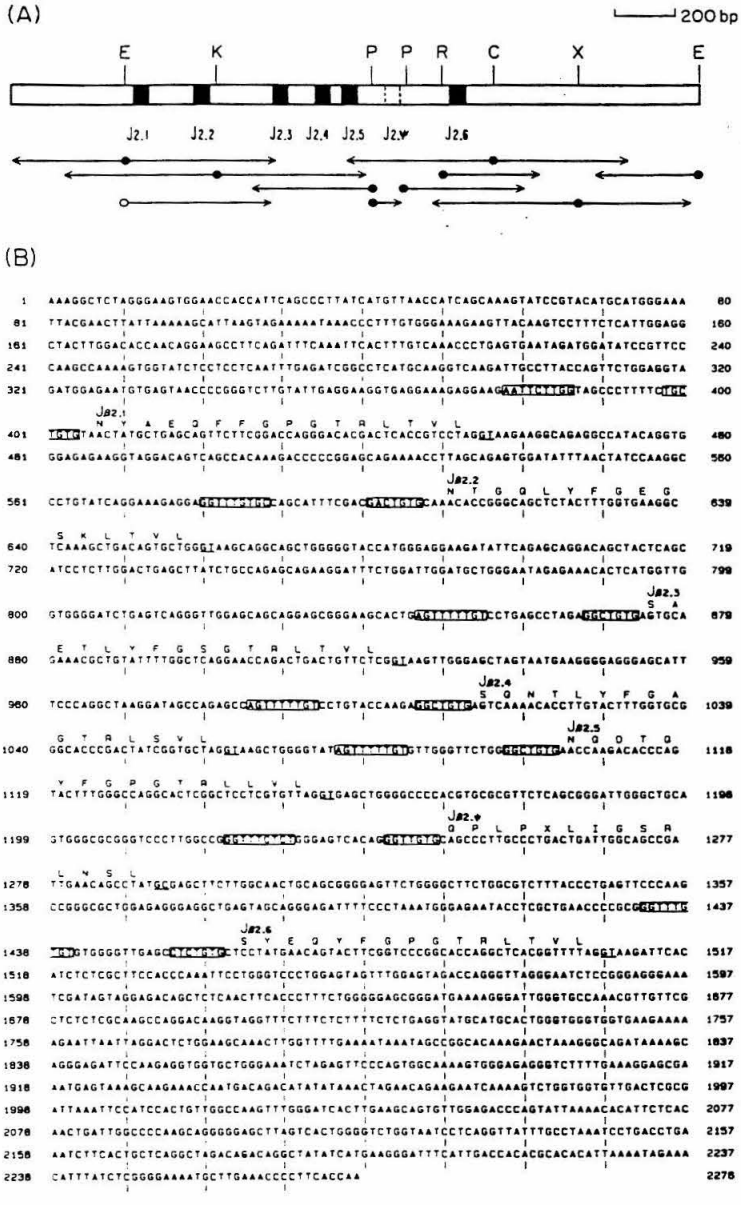


Figure 3. Chromosomal Organization of the $J_{\beta 2}$ Gene Cluster

(A) Partial restriction map of the $J_{\beta 2}$ gene cluster and the sequencing strategy employed to analyze this region. Solid boxes indicate functional $J_{\beta 2}$ gene segments ($J_{\beta 2.1}$ to $J_{\beta 2.6}$), and the dotted box indicates a possible pseudogene segment ($J_{\beta 2.\psi}$). Closed circles indicate a 3′ labeling and open circles a 5′ labeling for Maxam Gilbert sequence analysis. The restriction enzyme abbreviations are as follows: C, Cla I; E, Eco RI; K, Kpn I; P, Pst I; R, Sac II; X, Xba I.

(B) 2.3 kb of DNA sequence for the $J_{\beta 2}$ gene cluster and flanking sequence. The recognition sequences for DNA rearrangement have been boxed 5′ to each $J_{\beta 2}$ gene segment, and the RNA splice donor sites (GT) at their 3′ ends have been underlined. The predicted amino acid sequence encoded by each $J_{\beta 2}$ gene segment is shown above the DNA in the single-letter amino acid code (Dayhoff, 1978). The amino acid letters are placed above the second base of the codon. X indicates a termination codon.

tides to 49 nucleotides (Figure 4). At the DNA level they are 52% to 72% similar to each other. The translated sequences range in size from 15 to 16 amino acid residues and are 50% to 87% similar to each other (Figure 5). The N-terminal five residues of the $J_{\beta 2}$ segments are far more variable than their C-terminal regions.

In immunoglobulin genes, segments that are rearranged to form variable region genes have conserved DNA rearrangement recognition signal sequences adjacent to the sites of joining (Early et al., 1980). These recognition sequences are composed of three parts: a highly conserved heptameric sequence immediately adjacent to the V, D, or J gene segments; a nonconserved spacer sequence 11–12 or 22–23 nucleotides in length; and a relatively conserved nanomer sequence. The six $J_{\beta 2}$ gene segments have similar recognition sequences adjacent to their 5′ ends (Figure 4) and spacer sequences of 11–12 nucleotides in length. There do not appear to be unique recognition sequences for T cell receptor gene rearrangements (Figure 4).

At the 3′ side of the J gene segments are the canonical RNA splicing donor signals (Breathnach and Chambon,

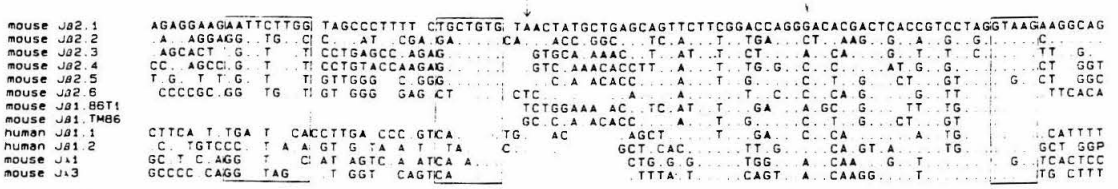


Figure 4. Comparison at the DNA Level of the Recognition Sequences for DNA Rearrangement, the Coding Region, and the 3′ Ends of Mouse and Human J_β Gene Segments and Two Mouse λ J Gene Segments

The mouse $J_{\beta2.1}$–$J_{\beta2.6}$ gene segments are from this paper. The $J_{\beta1.86T1}$ and $J_{\beta1.TM86}$ sequences are from two mouse cDNA clones (Hedrick et al., 1984b). The $J_{\beta1.1}$ and $J_{\beta1.2}$ sequences are from two linked human J_β gene segments (Siu et al., 1984). The $J_{\lambda1}$ and $J_{\lambda3}$ are two mouse λ joining gene segments (Blomberg and Tonegawa, 1982). The recognition sequences for rearrangement and the RNA donor splicing sites are boxed. The arrow indicates the beginning of the frame of translation. Dots indicate identity with the first sequence listed. Blanks indicate that gaps have been introduced during the alignment.

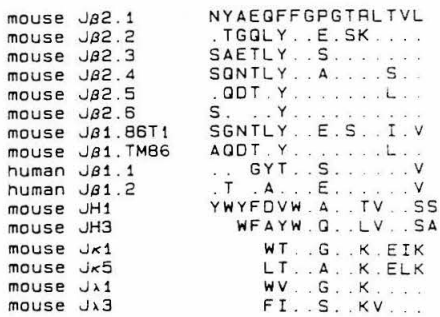


Figure 5. Comparison of the Translated Sequences of Mouse and Human J_β Segments and Various Immunoglobulin Light and Heavy Chain J Segments

The J_β gene segments are named as in Figure 4 with the addition of mouse $J_{\kappa1}$ and $J_{\kappa5}$ (Max et al., 1981) light chain joining segments and J_{H1} and J_{H3} (Sakano et al., 1980) heavy chain joining segments.

1981). Furthermore, the point of RNA splicing occurs between the first and second bases in the junctional codons, a conserved feature seen in all immunoglobulin, T cell receptor, and MHC genes studied to date, with a single exception to be discussed later.

It is interesting to note that a mouse β chain cDNA clone, TM86 (Hedrick et al., 1984b), contains a J coding segment that is identical with $J_{\beta2.5}$ except for the first two nucleotides. The differences are not silent and result in one amino acid difference upon translation. The J gene segment involved in this rearrangement could be encoded by a J_β gene segment, currently unidentified, from either the $J_{\beta1}$ or $J_{\beta2}$ gene clusters. A possibility we consider more likely is that the TM86 J sequence is encoded by $J_{\beta2.5}$, and the differences result from an imprecise joining of a D into $J_{\beta2.5}$, generating a novel codon. An analogous process is responsible for similar junctional diversity in immunoglobulins (Early and Hood, 1981), where such joining flexibility results in somatic expansion of the immunoglobulin repertoire.

A possible J_β pseudogene is located between the $J_{\beta2.5}$ and $J_{\beta2.6}$ gene segments. This region appears to be related to the functional J gene segments because its translated sequence exhibits homology to the other J_β gene segments (Figure 3) and it has appropriately placed 5′ rearrangement signal sequences. However, this gene segment has an in-frame termination codon, has lost the RNA splicing signal at the 3′ end, and has extensive differences in coding sequence from the other $J_{\beta2}$ gene segments. The translated sequence does not contain the sequence Gly-X-Gly found in all functional J gene segments (Figure 5).

The $C_{\beta2}$ Gene Has Four Exons

An analysis of the translated human and mouse β cDNA clones shows that the polypeptides can be divided into four distinct regions (Hedrick et al., 1984b; Yanagi et al., 1984): a region of about 113 residues that is homologous to immunoglobulin C region domains, including a disulfide bridge spanning 60 residues; 33 residues potentially forming a connecting peptide to the transmembrane region that is not homologous to any immunoglobulin regions; 22 relatively hydrophobic residues, probably encoding a transmembrane region; and five residues constituting a stop-transfer signal and a highly charged cytoplasmic region.

The complete nucleotide sequence of the $C_{\beta2}$ genomic gene segment, given in Figure 6, includes four coding regions determined by comparison with the cDNA sequences (Figure 7). The four exons of the $C_{\beta2}$ gene span about 1600 nucleotides and appear to be functional in that there are no in-frame termination codons and the RNA splicing signals for the exon boundaries are all canonical. Exon 1, 375 nucleotides in length, encodes the constant domain and a portion of the connecting peptide; exon 2, 18 nucleotides, encodes a short stretch of the connecting peptide; exon 3, 107 nucleotides, encodes the remainder of the connecting peptide and a probable transmembrane region; and exon 4 encodes one hydrophobic residue of the transmembrane region, the three lysine residues presumably constituting the stop-transfer signal associated with transmembrane regions, two additional hydrophilic residues corresponding to the cytoplasmic region, and about 160 nucleotides of the putative 3′ untranslated region including a canonical poly(A)$^+$ additional signal sequence. These exons are separated by introns of 506, 145, and 383 nucleotides, respectively.

The exons of the $C_{\beta2}$ gene do not correlate precisely

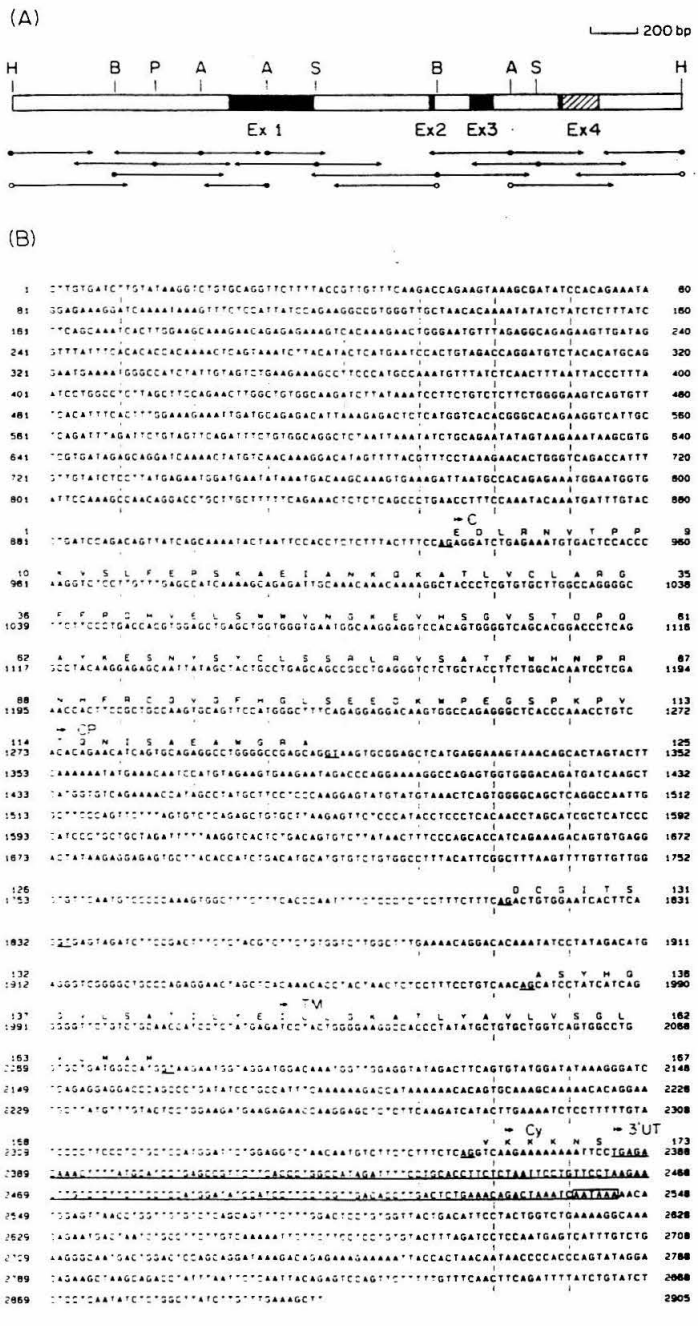


Figure 6. Organization and Sequence of the $C_{\alpha 2}$ Gene

(A) Partial restriction map of the $C_{\alpha 2}$ gene. Shaded areas indicate exons. The hatched area indicates the 3′ untranslated region. Open and closed circles are as described in Figure 3A. The restriction enzyme abbreviations are as follows: A, Ava II; B, Bgl II; H, Hind III; P, Pst I; S, Sac I.

(B) 2.9 kb DNA sequence containing the $C_{\alpha 2}$ gene and flanking sequences. The RNA splicing donor and acceptor sites are underlined as is the 3′ untranslated region. The predicted amino acid sequences encoded by the exons are shown above in the one-letter amino acid code. The following abbreviations are used: C, constant; CP, connecting peptide; TM, transmembrane; Cy, cytoplasmic; 3′UT, 3′ untranslated.

with the presumed functional regions or domains of the constant region. The connecting peptide is encoded on three separate exons. One of them, exon 2, encodes just six amino acids. Some immunoglobulins have a flexible proline-rich hinge peptide encoded by just such a small exon following the first constant region domain (Honjo, 1983). However, there is no sequence obviously similar to immunoglobulin hinge regions encoded by this exon or

Figure 7. Comparison at the DNA Level between the Coding Region of the $C_{\beta 2}$ Gene and the Constant Region Sequences of β Chain cDNAs from a Mouse T Cell (86T1) and a Human T Cell (YT35)

A vertical arrow above the sequences indicates an exon boundary. A horizontal arrow denotes the possible domain organization. The following abbreviations are used: C, constant; CP, connecting peptide; TM, transmembrane; Cy, cytoplasmic. 86T1: Hedrick et al., 1984b. YT35: Yanagi et al., 1984.

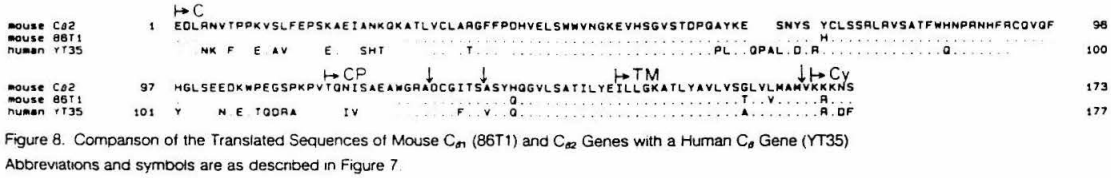


Figure 8. Comparison of the Translated Sequences of Mouse $C_{\beta 1}$ (86T1) and $C_{\beta 2}$ Genes with a Human C_{β} Gene (YT35)

Abbreviations and symbols are as described in Figure 7.

any other portion of the β gene. The third exon encodes both a portion of the connecting peptide and, along with one valine from exon 4, a stretch of 22 amino acids that is, apart from a single lysine, hydrophobic and uncharged and may encode a transmembrane segment. Bacterial rhodopsin has charged residues in its transmembrane segments, so presumably these charges can be shielded in the lipid bilayer (Kyte and Doolittle, 1982). If this lysine were excluded from the transmembrane region, the hydrophobic, uncharged sequence would only be 17 residues, far shorter than most known transmembrane regions (generally such hydrophobic segments range in length from 21 to 26 residues).

The nearly complete DNA sequence of a mouse cDNA clone 86T1 derived from the M12 T cell hybridoma has been published (Hedrick et al., 1984b). The cDNA clone appears to be encoded by the $C_{\beta 1}$ gene by two criteria. First, the 86T1 cDNA clone shares an Eco RI restriction enzyme site with the cosmid $C_{\beta 1}$ gene that is missing in the $C_{\beta 2}$ gene. Since this Eco RI restriction site is present in all inbred mouse strains tested (Caccia et al., 1984), we conclude that the differences between the 86T1 cDNA and $C_{\beta 2}$ DNA sequence cannot result from allelic differences between inbred strains. Second, this cDNA clone differs from the coding region of the $C_{\beta 2}$ gene by 23 nucleotides, of which five represent replacement substitutions (Figures 7 and 8). Four of the five replacement substitutions are in the 3′ coding region of the C_{β} genes. These differences are not likely to be due to somatic mutation since the same substitutions have been seen at the 3′ end of a second cDNA clone derived from the mouse T cell RBL5 tumor (data not presented). The coding regions of the $C_{\beta 1}$ and $C_{\beta 2}$ genes are closely related, being 96% similar at the DNA level and 97% similar at the protein level. These two genes may therefore have arisen by a relatively recent duplication. This supposition is consistent with the equidistant homology seen between both mouse C_{β} genes when compared with the human cDNA sequence (82%).

The RNA donor and acceptor splice signals for the four exons obey the established rules for these sites. RNA splicing occurs between the first and second bases in the codons at the exon boundaries between exons 1 and 2, as well as exons 2 and 3. RNA splicing between exons 3 and 4 occurs between codons (Figure 6B). This is to our knowledge the first exception to the rule that RNA splicing occurs between the first and second bases of the junctional exon codons in all of the gene families associated with the immune response (immunoglobulin, major histocompatibility, β_2-microglobulin, and T cell receptor genes).

The translated $C_{\beta 2}$ and $C_{\beta 1}$ genes have three possible sites of glycosylation. The glycosylation recognition sequences Asn-X-Ser or Asn-X-Thr occur at positions 5–7, 70–72, and 120–122 in the constant region (Figure 8). The human C_{β} gene shares only the second of these glycosylation sites and has no others.

In comparing the mouse $C_{\beta 1}$ and $C_{\beta 2}$ genes, the ratio of silent to replacement amino acid substitutions is approximately ten times what would be expected of random mutation (Miyata et al., 1980). This observation suggests

that a high degree of selective pressure maintains the coding sequences of these genes. The same skewed ratio of silent to replacement substitutions is also observed when either mouse C_α gene is compared with the human C_α gene. These observations are in keeping with the fact that the genes encoding T cell receptors have diverged far less than their immunoglobulin counterparts (Hedrick et al., 1984b; Honjo, 1983; Siu et al., 1984; Yanagi et al., 1984).

T Cells Can Rearrange Both $J_{\beta1}$ and $J_{\beta2}$ Gene Segments

The rearrangements of β genes seen in T cell lines appear to be more complex than rearrangements seen in B cells with immunoglobulin gene probes. We have obtained probes specific for the $J_{\beta1}$, $C_{\beta1}$, $J_{\beta2}$, and $C_{\beta2}$ gene-encoding regions, and these are indicated in Figure 2 as probes 1 through 4. The $C_{\beta1}$ and $C_{\beta2}$ probes hybridize more strongly to their corresponding gene segments than to each other, whereas the $J_{\beta1}$ and $J_{\beta2}$ probes do not cross-hybridize. In Figure 9A we have carried out Southern blot analyses on paired sets of liver DNA from the C57BL/6 mouse (germline) (left) and on DNA from a C57BL/6 helper cell clone L2 (right) (Glasebrook and Fitch, 1980), using these probes. Hybridization with the $C_{\beta1}$ probe (Pr2) indicates that both copies of the $C_{\beta1}$ gene are deleted in the L2 clone. Surprisingly, hybridization of the L2 clone with a probe 5′ of $C_{\beta1}$, $J_{\beta1}$ (Pr1), reveals a band that differs from the germ line. This indicates that either the two $J_{\beta1}$ homologs of the L2 clone have undergone a similar rearrangement or the L2 cells have rearranged the $J_{\beta1}$ homolog on one chromosome and deleted it on the other. Results with six other restriction enzymes (data not shown) support the second hypothesis. Only one band, non-germ-line, is found in all cases. Using a $J_{\beta2}$ probe (Pr3) we can see that the L2 clone has rearranged both alleles. The $C_{\beta2}$ genes are both present and in a germ-line configuration. Therefore, two rearrangements have occurred in the $J_{\beta2}$ cluster 5′ of the $C_{\beta2}$ gene (Figure 9A, lanes 5 and 6, 7 and 8). A schematic representation of the rearrangements occurring in L2 is shown in Figure 9B. Figure 10 shows the rearrangements observed in a T lymphoma cell line (BAL 3); both alleles of the $C_{\beta1}$ gene are maintained and two rearrangements have occurred 5′ to the $C_{\beta1}$ gene involving the $J_{\beta1}$ cluster.

Two important points emerge from these analyses. First, in individual T cells both $J_{\beta1}$ and $J_{\beta2}$ gene segments can participate in DNA rearrangements. Second, the DNA rearrangements in T cells appear to be somewhat more complex than their immunoglobulin counterparts. Multiple DNA rearrangement or deletional events may occur on a single chromosome, and rearrangements generally occur on both chromosomes. DNA rearrangements occur in all functional T cell helper and killer cells that have been tested (Kronenberg et al., unpublished data).

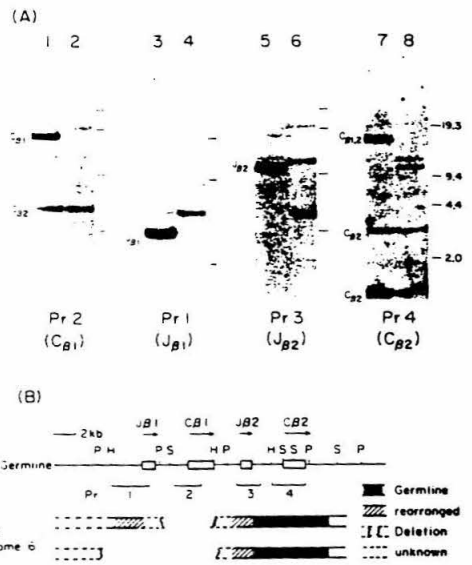


Figure 9. Southern Blot Analyses of T Cell DNA Illustrating J_β Gene Segment Rearrangements

(A) Southern blot analyses of liver DNA from C57BL/6 mice (lanes 1, 3, 5, and 7) and DNA from the T helper cell line L2 (lanes 2, 4, 6, and 9). The restriction enzymes used are as follows: Hind III for lanes 1, 2, 5, and 6; Pvu II for lanes 3 and 4; Sac I for lanes 7 and 8. Probes 1–4 are as shown in Figure 2. Size marker lengths are indicated in kilobases for one blot.
(B) Diagrammatic representation of the DNA rearrangement events that have presumably occurred in the L2 T cell line. The open boxes represent the approximate locations of the $J_{\beta1}$ and $J_{\beta2}$ gene segments and $C_{\beta1}$ and $C_{\beta2}$ genes. Restriction enzyme abbreviations: H, Hind III; P, Pvu II; S, Sac I.

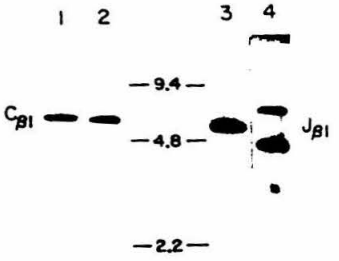


Figure 10. Southern Blot Analyses of T Cell DNA Illustrating $J_{\beta1}$ Gene Segment Rearrangements

DNA from BALB/c mouse liver (lanes 1 and 3) and from the BALB/c T lymphoma cell line BAL 3 (lanes 2 and 4) digested by Pvu II and hybridized with the indicated probes.

Discussion

We have isolated a cosmid clone containing mouse germline J_β gene segments and C_β genes and have used this clone to determine some features of the structure and organization of the T cell receptor β gene locus. Several conclusions can be drawn from our analysis of the germline β genes. First, there are two closely related C_β genes separated by 6 kb and in the same transcriptional orientation. Comparison of the coding region of the $C_{\beta 2}$ gene with the previously published mouse C_β cDNA sequence (Hedrick et al., 1984b) indicates that these two C genes are 96% similar at the nucleotide level. Second, the $C_{\beta 2}$ gene is split into four exons. The exon–intron boundaries do not correspond obviously to recognizable domains of the translated protein sequence. Third, each C_β gene is preceded by a cluster of J_β gene segments also in the same transcriptional orientation. The nucleotide sequences of these gene segments show significant homology to immunoglobulin J gene segments, including the presumed rearrangement signal sequences found 5′ to all immunoglobulin J gene segments.

The data presented here strengthen the conclusion that the T cell receptor β chain gene family is related to the immunoglobulin gene families. Figure 11 compares the organization of the β chain genes to the three antibody gene families. Are the T cell receptor β chain genes more evolutionarily related to the light chain gene family or to the heavy chain gene family? The β chain genes share three features with the heavy chain family that are distinct from the light chain gene families. First, there is strong evidence for the existence of D_β gene segments (Siu et al., 1984). However, these gene segments have not been isolated. Second, the C_β genes contain a sequence encoding a probable hydrophobic transmembrane region. The β polypeptides, like immunoglobulin heavy chains, are therefore likely to be integral membrane proteins. Third, the $J_{\beta 2}$ gene segments encode 15–16 amino acids. This length is similar to that of the J_H gene segments, but longer than the J gene segments of either light chain (Figure 5). The difference in length is at the 5′ end of the J gene segments, the region encoding the most divergent portion of both J_β and J_H regions. In immunoglobulin heavy chains this sequence encodes part of the third hypervariable region and therefore is probably involved in antigen binding.

Although there are important similarities between the T cell β genes and heavy chain genes, it is not conclusive that the β genes are more closely related to these than to immunoglobulin light chain genes. When aligned and compared with the constant region domains of most known immunoglobulin light and heavy chains, the C or immunoglobulin-like region of the T cell β chain is most consistently similar to λ light chains (about 34%–37%). The next most similar comparison is to the first domain of the various γ heavy chains (about 32%–36%). Most all other comparisons, including κ light chains, show less than 30% similarity. In addition, the organization and structure of the J_β gene segments share distinctive features with those of immunoglobulin λ genes. The spacer region of the rearrangement signal sequence is about 12 nucleotides long for both λ and β J gene segments, but 22 for κ and heavy chains. Lambda constant region genes also are closely linked, each with its own associated J gene segment (Figure 11). If the T cell receptor β genes diverged from immunoglobulin genes during or prior to the divergence of the immunoglobulin light and heavy chain genes, we might expect the β genes to be equally related to the three immunoglobulin gene families.

A unique feature of the T cell β genes is the presence of multiple J gene segments in front of each of the C_β genes. Both clusters of J_β gene segments probably rearrange with the same families of V_β gene segments. The C_μ gene of the immunoglobulin heavy chain gene family is most 3′-proximal to the J_H gene segments. The C_H genes farther downstream from the C_μ gene can be expressed in either of two ways. They can be transcribed with the C_μ gene on a relatively long primary transcript that also includes a complete variable region gene. Splicing of the C_H gene in question to J_H can then form a complete heavy

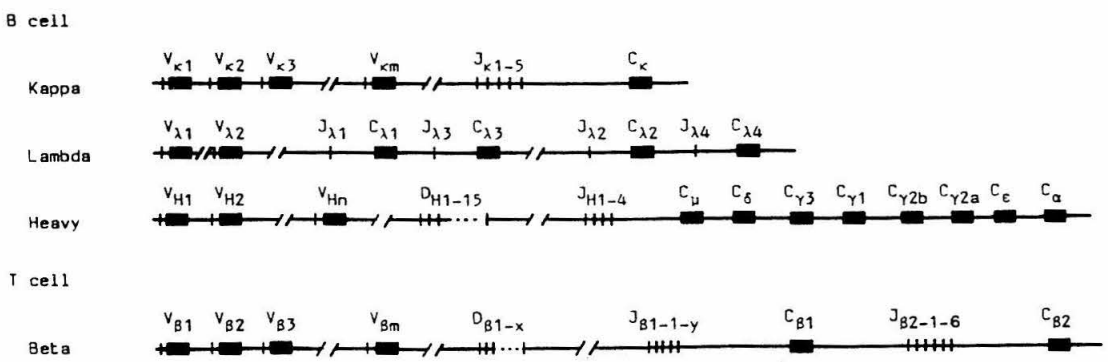


Figure 11. Diagrammatic Representation of the Organization of Immunoglobulin and T Cell Receptor Genes
Breaks in the horizontal lines indicate sequences that have not yet been physically linked.

chain messenger RNA. This mechanism is best characterized for the C_δ gene located about 4 kb 3′ to the C_μ gene (Moore et al., 1981). Or they can replace the C_μ gene by directly rearranging near the J_H gene segment with the loss of all other intervening DNA and C genes (Davis et al., 1980). This DNA rearrangement or class switching is mediated via sequences 5′ to most C_H genes that contain multiple, short repeats (switch sequences). It remains to be determined whether the $C_{\beta 2}$ gene is expressed by either of these mechanisms. The presence of J gene segments in front of the $C_{\beta 2}$ gene may obviate the need for either a long primary transcript or switch recombination, as seems to be the case for λ chain expression.

Finally, we might ask whether there is any functional significance to the presence of these two C_β genes in the mouse genome. Clearly both C_β genes can be expressed. In addition, the data presented here indicate that in T lymphocytes DNA rearrangements occur to both J_β gene clusters. The close sequence similarity of the $C_{\beta 1}$ and $C_{\beta 2}$ genes is most likely the result of a recent gene duplication event. It seems unlikely that the few sequence differences between the $C_{\beta 1}$ and $C_{\beta 2}$ genes are of great biological significance.

Experimental Procedures

Construction and Screening of the Cosmid Library

A cosmid library was constructed from Mbo I partially digested B10.WR7 liver DNA and the cosmid vector pTL5 (Lund et al., 1982) according to Grosveld et al. (1981) and Steinmetz et al. (1982a). The library contains about 600,000 colonies distributed on 60 filters.

Screening of the cosmid library was carried out according to Steinmetz et al. (1982b). The hybridization was for 18 hr at 65°C with a probe concentration of 10^5 cpm/ml (the specific activity of the probe was about $1–3 \times 10^8$ cpm/μg). The filters were then washed five times for 30 min each at 65°C in 2× SSC, 0.1% SDS.

Hybridization Probes

To obtain hybridization probes from the cosmid clone 2.3W7, a Southern blot of the cloned DNA was hybridized to total mouse DNA to a Cot that only allows detection of repetitive sequences (Steinmetz et al., 1980). The fragments that also hybridized to human T cell cDNA sequences were isolated and used as probes against mouse liver DNA in the presence of 25 μg/ml total mouse DNA as competitor. The single- or low-copy fragments were subcloned in pUC8 (Vieira and Messing, 1982). These are indicated in Figure 2.

Southern Blots

Ten micrograms of restriction enzyme digested liver or T cell DNA was separated by electrophoresis on a 0.6% or 0.8% agarose gel and transferred to nitrocellulose (Southern, 1975). The hybridizations were performed under the same conditions as for the cosmid screening, but with a probe concentration of 10^6 cpm/ml.

DNA Sequencing

DNA sequencing was done by the method of Maxam and Gilbert (1980). The 3′-end labeling was done either by filling in a recessed end using the large fragment of DNA polymerase I or by using deoxynucleotidyl terminal transferase to add [^{32}P]-α-dideoxy ATP to an overhanging 3′ end. The 5′-end labeling was done using [^{32}P]-γ-ATP and polynucleotide kinase after treatment with calf intestinal phosphatase.

Acknowledgments

The authors thank Elizabeth Gibb for her help during the construction of the cosmid library, Dr. Suzanna Horvath for the synthetic oligonucleotides, Gerald Siu and Drs. Martha Zuniga, Joan Kobori, and Richard Barth for critical reading of the manuscript. We also want to thank Debbie Maloney for technical assistance and Bertha Jones, Bernita Larsh, and Connie Katz for the preparation of the manuscript. M. M. is supported by the Centre National de la Recherche Scientifique (France).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received May 14, 1984

References

Acuto, O., Hussey, R., Fitzgerald, K., Protentis, J., Meuer, S., Schlossman, S., and Reinherz, E. (1983). The human T cell receptor: appearance in ontogeny and biochemical relationship of α and β subunits on IL-2 dependent clones and T cell tumors. Cell *34*, 717–726.

Bernard, O., Hozumi, N., and Tonegawa, S. (1978). Sequences of mouse immunoglobulin light chain genes before and after somatic changes. Cell *15*, 1133–1144.

Blomberg, B., and Tonegawa, S. (1982). DNA sequences of the joining regions of mouse λ light chain immunoglobulin genes. Proc. Nat. Acad. Sci. USA *79*, 530–533.

Breathnach, R., and Chambon, P. (1981). Organization and expression of eucaryotic split genes coding for proteins. Ann. Rev. Biochem. *50*, 349–383.

Caccia, N., Kronenberg, M., Saxe, D., Haars, R., Bruns, G., Goverman, J., Malissen, M., Willard, H., Simon, M., Hood, L., and Mak, T. W. (1984). The β T cell receptor genes are located on chromosome 7 in humans and chromosome 6 in mice. Cell *37*, 1091–1099.

Davis, M. M., Kim, S. K., and Hood, L. (1980). DNA sequences mediating class switching in α-immunoglobulins. Science *209*, 1360–1365.

Dayhoff, M. O. (1978). Atlas of Protein Sequence and Structure, *5*, Suppl. 2, (Washington, D.C.: Nat. Biomed. Res. Found.).

Early, P., and Hood, L. (1981). Allelic exclusion and nonproductive immunoglobulin gene rearrangements. Cell *24*, 1–3.

Early, P., Huang, H., Davis, M., Calame, K., and Hood, L. (1980). An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: V_H, D and J_H. Cell *19*, 981–992.

Glasebrook, A. L., and Fitch, F. W. (1980). Alloreactive cloned T cell lines. I. Interactions between cloned amplifier and cytolytic T cell lines. J. Exp. Med. *151*, 876–895.

Grosveld, F. G., Dahl, H. H. M., de Boer, E., and Flavell, R. A. (1981). Isolation of β globin-related genes from a human cosmid library. Gene *13*, 227–237.

Hedrick, S., Cohen, D., Nielsen, E., and Davis, M. (1984a). Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. Nature *308*, 149–153.

Hedrick, S., Nielsen, E., Kavaler, J., Cohen, D., and Davis, M. (1984b). Sequence relationships between putative T-cell receptor polypeptides and immunoglobulins. Nature *308*, 153–158.

Honjo, T. (1983). Immunoglobulin genes. Ann. Rev. Immunol. *1*, 499–528.

Kappler, J., Kubo, R., Haskins, K., White, J., and Marrack, P. (1983). The mouse T cell receptor: comparison of MHC-restricted receptors on two T cell hybridomas. Cell *34*, 727–737.

Kronenberg, M., Kraig, E., and Hood, L. (1983). Finding the T-cell antigen receptor: past attempts and future promise. Cell *34*, 327–329.

Kyte, J., and Doolittle, R. F. (1982). A simple method of displaying the hydropathic character of a protein. J. Mol. Biol. *157*, 105–132.

Lund, T., Grosveld, F. G., and Flavell, R. A. (1982). Isolation of transforming DNA by cosmid rescue. Proc. Nat. Acad. Sci. USA *79*, 520–524.

Max, E. E., Maizel, J. V., and Leder, P. (1981). The nucleotide sequence of a 5.5-kilobase DNA segment containing the mouse κ immunoglobulin J and C region genes. J. Biol. Chem. *256*, 5116–5120.

Maxam, A., and Gilbert, W. (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. Meth. Enzymol. *65*, 499–560.

McIntyre, B., and Allison, J. (1983). The mouse T cell receptor: structural heterogeneity of molecules of normal T cells defined by xenoantiserum. Cell *34*, 739–746.

Miyata, T., Yasunaga, T., and Nishida, T. (1980). Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc. Nat. Acad. Sci. USA *77*, 7328–7332.

Moore, K. W., Rogers, J., Hunkapiller, T., Early, P., Nottenburg, C., Weissman, N. I., Bazin, H., Wall, R., and Hood, L. E. (1981). Expression of IgD may use both DNA rearrangement and RNA splicing mechanisms. Proc. Nat. Acad. Sci. USA *78*, 1800–1804.

Sakano, H., Maki, R., Kurosawa, Y., Roeder, W., and Tonegawa, S. (1980). Two types of somatic recombination are necessary for the generation of complete immunoglobulin heavy-chain genes. Nature *286*, 676–683.

Samelson, L., Germain, R., and Schwartz, R. (1983). Monoclonal antibodies against the antigen receptor on a cloned T-cell hybrid. Proc. Nat. Acad. Sci. USA *80*, 6972–6976.

Siu, G., Clark, S. P., Yoshikai, Y., Malissen, M., Yanagi, Y., Strauss, E., Mak, T. W., and Hood, L. (1984). The human T-cell antigen receptor is encoded by variable, diversity and joining gene segments that rearrange to generate a complete V gene. Cell *37*, 393–401.

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. *98*, 503–517.

Steinmetz, M., Hochtl, J., Schnell, H., Gebhard, W., and Zauchau, H. G. (1980). Cloning of V region fragments from mouse liver DNA and localization of repetitive DNA sequences in the vicinity of immunoglobulin gene segments. Nucl. Acids Res. *8*, 1721–1729.

Steinmetz, M., Winoto, A., Minard, K., and Hood, L. (1982a). Clusters of genes encoding mouse transplantation antigens. Cell *28*, 489–498.

Steinmetz, M., Minard, K., Horvath, S., McNicholas, J., Frelinger, J., Wake, C., Long, E., Mach, B., and Hood, L. (1982b). A molecular map of the immune response region from the major histocompatibility complex of the mouse. Nature *300*, 35–42.

Tonegawa, S. (1983). Somatic generation of antibody diversity. Nature *302*, 575–581.

Vieira, J., and Messing, J. (1982). The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. Gene *19*, 259–268.

Yanagi, Y., Yoshikai, Y., Leggett, K., Clark, S., Aleksander, I., and Mak, T. (1984). A human T cell-specific cDNA clone encodes a protein having extensive homology to immunoglobulin chains. Nature *308*, 145–149.

Note Added in Proof

One Pvu II site has been omitted in Figure 2 and Figure 9. It is 19.2 kb from the 5′ end of the cosmid clone.

Cell, Vol. 40, 859–867 April 1985. Copyright © 1985 by MIT

0092-8674/85/040859-09 $02.00/0

Rearranged β T Cell Receptor Genes in a Helper T Cell Clone Specific for Lysozyme: No Correlation between V_β and MHC Restriction

Joan Goverman,* Karyl Minard,* Nilabh Shastri,* Tim Hunkapiller,* Dan Hansburg,† Eli Sercarz,‡ and Leroy Hood*

* Division of Biology
California Institute of Technology
Pasadena, California 91125
† Department of Pathology
Fox Chase Cancer Center
Philadelphia, Pennsylvania 19111
‡ Department of Microbiology
University of California, Los Angeles
Los Angeles, California 90024

Summary

The helper T cell clone 3H.25 is specific for hen egg white lysozyme and the class II MHC molecule I-A^b. This T_H cell has three rearrangements in the β-chain gene family—a $V_{\beta 1}$-$D_{\beta 1}$-$J_{\beta 1}$, and a $D_{\beta 2}$-$J_{\beta 2}$ rearrangement on one homolog and a $D_{\beta 1}$-$J_{\beta 2}$ rearrangement on the other. These observations demonstrate that this functional T lymphocyte expresses only a single V_β gene segment and, accordingly, exhibits allelic exclusion of β-chain gene expression. The rearranged 3H.25 V_β gene segment is the same as that expressed in a T helper cell specific for cytochrome c and an I-E^k MHC molecule. Thus, there is no simple correlation between the V_β gene segment and antigen specificity or MHC restriction.

Introduction

B and T cells recognize antigen through receptors expressed on their cell surfaces—the immunoglobulin molecules on B cells and the antigen-specific receptors on T cells. The extensive diversity of receptor repertoires exhibited by both types of lymphocytes results from similar strategies of gene organization and genetic diversification (Early and Hood, 1981a; Honjo, 1983; Tonegawa, 1983). Both receptors are disulfide-linked heterodimers of light and heavy chains for immunoglobulins or α and β chains for T cell receptors. Each chain is composed of a variable region involved with binding antigen and a constant region that defines the type or class of the chain. Genes encoding the variable regions of immunoglobulins are formed by DNA rearrangement events that juxtapose either two gene segments—variable (V_L) and joining (J_L)—in light chains (Bernard et al., 1978) or three gene segments—V_H, diversity (D_H), and J_H—in heavy chains (Early et al., 1980; Sakano et al., 1980). The constant gene (C) is separated from the J gene segment by an intron that is removed during RNA processing. The T cell β chain also has a variable region encoded by three distinct gene segments, V_β, D_β, and J_β, which rearrange during T cell differentiation to generate a complete V_β gene (Siu et al., 1984a; Chien et al., 1984). Rearrangement of both immunoglobulin and β-chain receptor gene segments appears to be mediated by similar DNA recognition sequences.

The β-chain gene family is located on chromosome 6 of the mouse (Caccia et al., 1984; Lee et al., 1984) and has two closely linked C_β genes, $C_{\beta 1}$ and $C_{\beta 2}$, each associated with a cluster of six functional J_β gene segments about 2 kb upstream (Gascoigne et al., 1984; Malissen et al., 1984). One D_β gene segment, $D_{\beta 2.1}$, has been identified 578 bases to the 5′ side of the $J_{\beta 2}$ gene cluster. Another D_β gene segment, $D_{\beta 1.1}$, is located 647 bases 5′ to the $J_{\beta 1}$ gene cluster (Siu et al., 1984b; Kavaler et al., 1984). An unknown number of additional $D_{\beta 1}$ gene segments may lie further upstream. Because each C_β gene has its own cluster of J_β gene segments, V_β gene segments may rearrange directly to either C_β gene. Two types of β-chain gene rearrangements have been described, V_β-D_β-J_β joinings that generate a complete V_β gene and D_β-J_β only, or incomplete rearrangements in which no complete V_β gene is produced (Siu et al., 1984b; Kavaler et al., 1984; Kronenberg et al., 1985). These rearranged β-chain genes can be expressed as 1.3 kb and 1.0 kb RNA transcripts, respectively (Clark et al., 1984; Siu et al., 1984b). In both immunoglobulins and the β chain of T cell receptors, the diversity of the variable regions arises from at least three shared strategies: a multiplicity of germ-line gene segments, combinatorial joining of the gene segments, and somatic variation associated with the joining events.

Despite similarities in the ways in which B and T cells generate V-region diversity, T cells differ from their B cell counterparts in several striking regards. While B cells can bind to free antigen, T cells recognize antigen only when it is presented on a cell surface in association with specific polypeptides encoded by the major histocompatibility complex (MHC), a property referred to as MHC restriction (Golub, 1980; Rosenthal and Shevach, 1973; Schwartz et al., 1978; Katz et al., 1973). There are three recognized subsets of T lymphocytes: cytotoxic T cells (T_C), which kill target cells expressing antigen on the cell surface, and helper (T_H) and suppressor (T_S) T cells, which stimulate or suppress immune responses, respectively. Antigen recognition by each subset is restricted by a different MHC molecule: cytotoxic T cells recognize antigen in association with class I MHC gene products (Doherty and Zinkernagel, 1975) and helper T cells recognize antigen in association with Class II I-A or I-E gene products (Thomas et al., 1977). The nature of MHC restriction in suppressor T cells is less clear (Germain and Benacerraf, 1981). The molecular basis for the simultaneous recognition by a T cell of antigen and MHC molecules is not yet understood. A variety of experiments suggest that recognition of the antigen and the MHC molecule is mediated by a single T cell receptor molecule rather than by two receptors, one specific for each (Kappler et al., 1981; Heber-Katz et al., 1982; Hunig and Bevan, 1982). However, this latter possibility has not been eliminated. Either the $C_{\beta 1}$ or the $C_{\beta 2}$

gene may be employed by helper and cytotoxic T cells, whereas suppressor T cells often appear to delete both C_β genes (Kronenberg et al., 1985).

We are interested in the mechanisms responsible for generating T cell receptor diversity and the relationship between rearrangement of T cell receptor genes and their expression on the cell surface. In B cells the expression of immunoglobulins follows a principle known as allelic exclusion; that is, only one functional light chain and one functional heavy chain in each B cell is associated to produce a functional immunoglobulin molecule. Allelic exclusion occurs because only one light chain gene and one heavy chain gene are productively rearranged—the other alleles are not rearranged or are nonproductively rearranged. Nonproductive rearrangements result either from D-J joinings without a V-gene segment or from V-D-J joinings that cannot be translated into a polypeptide that will associate with its counterpart chain (Bernard et al., 1981; Kwan et al., 1981; Early and Hood, 1981b; Alt et al., 1982). The question of whether the principle of allelic exclusion operates in T cells and what mechanisms may function in generating receptor diversity can be approached by analyzing all of the β-chain gene rearrangements that occur in functional T cells. We report here an analysis of all of the β-chain gene rearrangements that have occurred in the cloned helper T cell line 3H.25. This T_H clone was generated from inbred C57BL/6 mice and is restricted by the I-A^b molecule (superscript b denotes the MHC haplotype of the C57BL/6 mouse). This clone is specific for the antigen hen egg white lysozyme (HEL) and requires this antigen and syngeneic antigen-presenting cells for growth in culture. Three J_β gene rearrangements have been identified by Southern blot analyses with appropriate β-chain probes. These rearranged sequences were cloned and analyzed. The T_H 3H.25 cell exhibits allelic exclusion of its β-chain genes in that only one of the β-chain gene rearrangements is functional. The functionally rearranged V_β gene segment in T_H 3H.25 is the same as that employed in the T_H clone 2B4, which is specific for the antigen cytochrome C. This observation is discussed in terms of the relationship of the V_α gene segment to antigen and/or MHC specificity and the size of the V_β gene repertoire.

Results and Discussion

Three β-Chain Rearrangements Have Occurred in T_H 3H.25

T_H 3H.25 and germ-line or unrearranged DNAs were analyzed by Southern blot analyses using β-chain probes specific for the $D_{\alpha1.1}$, $J_{\alpha1}$, and $J_{\alpha2}$ gene segments and for the $C_{\alpha1}$ and $C_{\alpha2}$ genes (Figure 1). When T cell and liver (germ-line) DNAs are digested with the restriction enzyme Bam HI and analyzed by Southern blot analyses with a $C_{\beta1}$ probe, two germ-line bands, $C_{\alpha1}$ and $C_{\alpha2}$, are observed in the liver DNA, and one germ-line, $C_{\alpha2}$, and one rearranged band, $C_{\alpha1}$, are noted in the T cell DNA (Figure 2, lanes 1 and 2). Because the Bam HI sites flanking $C_{\alpha1}$ lie 5′ to some of the $J_{\alpha1}$ gene segments and 3′ to the $J_{\alpha2}$ gene segments, rearrangement events occurring on either

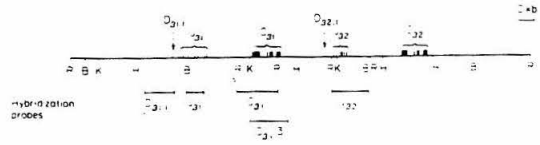


Figure 1. Partial Restriction Map of Germ-line T Cell Receptor β Genes

Vertical lines or boxes above the horizontal line indicate the exons encoding $D_{\alpha1.1}$, $D_{\alpha2.1}$, J_α, and C_α gene segments. Restriction enzyme abbreviations are as follows: B, Bam HI; H, Hind III; K, Kpn I; R, Eco RI. Cleavage sites are indicated by vertical lines below the horizontal line. Hybridization probes were purified restriction fragments from subclones made from the germ-line β-gene cosmid 2.3W7 (Malissen et al., 1984) or from a $C_{\beta1}$ cDNA clone. The $D_{\beta1.1}$ probe is a 1.8 kb Pst I fragment that has been subcloned into M13 mp8 and detects sequences 5′ to $D_{\beta1.1}$ in germ-line DNA. The $J_{\beta1}$ probe is a Bam HI–Pvu II fragment isolated from a 6.2 kb Bam HI–Hind III fragment subcloned into pUC8. It hybridizes to $J_{\beta1.3}$, $J_{\beta1.4}$, and $J_{\beta1.5}$. The $C_{\beta1}$ probe is a 2.5 kb Eco RI fragment derived from the same subclone as the $J_{\beta1}$ probe. The $C_{\beta1}$ 3′ probe is a 700 bp Pst I fragment of the RBL5-17 cDNA clone (Caccia et al., 1984) that hybridizes to most of the constant region and 3′ untranslated sequences of $C_{\beta1}$. The $J_{\beta2}$ probe contains all of the $J_{\beta2}$ gene segments and is a 2.3 kb Eco RI fragment isolated from a 4.5 kb Bam HI–Hind III fragment subcloned into pUC8. The $C_{\beta1}$ and $C_{\beta1}$ 3′ probes also hybridize to $C_{\beta2}$ as well as $C_{\beta1}$ sequences because of homology between these two genes.

side of $C_{\beta1}$ in the $J_{\beta1}$ or $J_{\beta2}$ gene clusters (or both) cannot be distinguished with this enzyme. To determine whether a rearrangement of $D_{\beta2.1}$ to a $J_{\beta2}$ gene segment had occurred, Eco RI digests of liver and T_H 3H.25 DNA were hybridized with a 3′ $C_{\beta1}$ probe. This probe detects rearrangement of the Eco RI fragment containing the $D_{\beta2.1}$ gene segment (Figure 1). Lane 4 of Figure 2 shows that this fragment, which is 3.2 kb in germ-line DNA, is rearranged to a 4.1 kb fragment consistent with a deletion of the Eco RI site between the $D_{\beta2.1}$ and $J_{\beta2}$ gene segments. This demonstrates that a $D_{\beta2.1}$-$J_{\beta2}$ rearrangement is present in T_H 3H.25 cells. One rearranged band corresponding to 3′ $C_{\beta1}$ sequences indicates again that only one copy of $C_{\beta1}$ is present in this cell line.

Rearrangements in the $J_{\beta1}$ gene cluster were examined using the $D_{\beta1}$ and $J_{\beta1}$ probes (Figure 1). Southern blots of Kpn- and Eco RI-digested T_H 3H.25 and liver DNAs hybridized with the $J_{\beta1}$ probe detected one rearranged $J_{\alpha1}$ band in the T_H 3H.25 DNA, consistent with a single rearranged $J_{\beta1}$-$C_{\beta1}$ fragment (Figure 2, lanes 5–8). A Southern blot analysis of Eco RI-digested T cell and liver DNAs with the $D_{\beta1.1}$ probe (Figure 2, lanes 9 and 10) revealed a single rearranged band of 7.1 kb, which is distinct from the 4.1 kb $J_{\beta1}$ band seen in a similar digest (Figure 2, lane 7). This probe will only detect sequences that lie 5′ to the $D_{\beta1.1}$ gene segment in germ-line DNA. This implies that there is a single rearrangement of the $D_{\beta1.1}$ gene segment, which has not joined to a V_β gene segment in the T_H 3H.25 clone, and that this rearrangement is distinct from that of the $J_{\beta1}$ gene cluster. A Kpn Southern blot hybridized with the $C_{\beta1}$ probe establishes that the $J_{\alpha1}$ rearrangement and $C_{\beta1}$ gene are on the same DNA fragment (data not shown). The simplest explanation for these observations is that a $D_{\beta1}$ gene segment has joined to the $J_{\beta1}$ gene cluster on one chromosome while there has

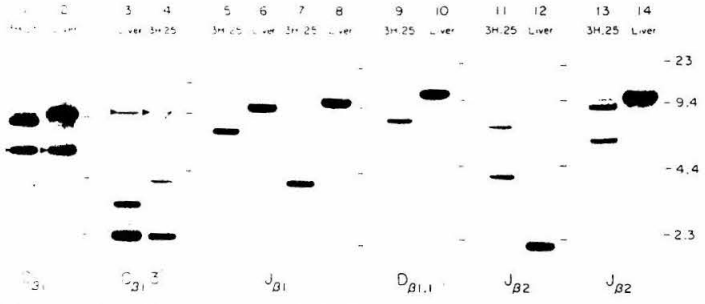


Figure 2. β-Gene Rearrangements in T_H 3H.25

T_H 3H.25 and liver DNAs were digested with: lanes 1 and 2, Bam HI; lanes 3 and 4, Eco RI; lanes 5 and 6, Kpn I; lanes 7–12, Eco RI; lanes 13 and 14, Bam HI. The probes are described in Figure 1. In lanes 1–4, the $C_{\beta 2}$ bands detected by cross-hybridization are indicated by carets. The migration distances of molecular weight markers are indicated.

been a $D_{\beta 1.1}$ rearrangement, presumably to the $J_{\beta 2}$ gene cluster on the second, resulting in the deletion of the $J_{\beta 1}$ cluster and the $C_{\beta 1}$ gene from this chromosome.

To examine rearrangements in the $J_{\beta 2}$ gene cluster, Southern blots of Eco RI- and Bam HI-digested T_H 3H.25 and liver DNAs were examined with the $J_{\beta 2}$ probe (Figure 2, lanes 11–14). Two distinct rearranged bands were noted in each case, suggesting that both $J_{\beta 2}$ gene clusters are rearranged.

Together, these data demonstrate that there are three β-chain rearrangements in T_H 3H.25 DNA: a V_β-$D_{\beta 1}$-$J_{\beta 1}$ and a $D_{\beta 2.1}$-$J_{\beta 2}$ rearrangement on one chromosome and a $D_{\beta 1.1}$-$J_{\beta 2}$ rearrangement on the second. This latter rearrangement results in the deletion of the intervening sequence including the $C_{\beta 1}$ gene and $D_{\beta 2.1}$ gene segment.

T_H 3H.25 Has Only One Functional β Rearrangement

A complete genomic library was constructed from T_H 3H.25 DNA and the rearranged clones were selected using $J_{\beta 1}$ and $J_{\beta 2}$ probes. Restriction enzyme analyses identified two classes of clones, represented by 3H1.74 and 3H2.22 (Figure 3). The positions within the clones of the D_β and J_β gene segments and the C_β genes were determined using the probes indicated in Figure 1. The restriction fragments contained in these clones account for all the rearrangements detected in the Southern blot analyses of T_H 3H.25 DNA. The regions of the clones corresponding to the rearranged gene segments were subcloned and sequenced by the strategies outlined in Figure 3. The DNA sequences are given in Figure 4.

These sequences establish that the T_H 3H.25 clone has a complete V_β-$D_{\beta 1.1}$-$J_{\beta 1.2}$ and an incomplete $D_{\beta 2.1}$-$J_{\beta 2.5}$ rearrangement on one chromosome and an incomplete $D_{\beta 1.1}$-$J_{\beta 2.5}$ rearrangement on the second chromosome. The rearranged V_β gene has a leader sequence and an open reading frame throughout and presumably encodes a functional β-chain V region. Both incompletely rearranged D_β gene segments retain germ-line sequences on their 5′ sides.

A probe for the rearranged V_β gene in T_H 3H.25 was prepared from a 250 bp Eco RI–Rsa I fragment contained entirely within the V_β gene segment. A Northern blot analysis was carried out on poly(A)$^+$ selected RNA from T_H 3H.25 cells (Figure 5, lane 1). This V_β probe hybridizes to a 1.3 kb transcript, which also hybridizes to a $C_{\beta 1}$-specific gene probe derived from 3′ nontranslated $C_{\beta 1}$ sequences (Figure 5, lane 3) (Kronenberg et al., 1985). Thus a functional 1.3 kb RNA transcript is present in the T_H 3H.25 cells.

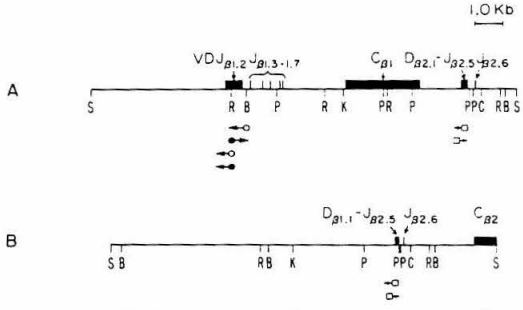


Figure 3. Organization of Genomic Clones Containing the β-Gene Rearrangements in T_H 3H.25

(A) Partial restriction map of 3H1.74 and the sequencing strategy employed to analyze this clone. Exons encoding V_β, D_β, J_β, and C_β gene segments are indicated by vertical lines or boxes above the horizontal line. Restriction enzyme abbreviations are as follows: B, Bam HI; C, Cla I; K, Kpn I; P, Pst I; R, Eco RI; S, Sal I. The VDJ rearranged region was sequenced by the procedure of Maxam and Gilbert with closed circles indicating a 3′-end-labeling and open circles a 5′-end-labeling. A 2.0 kb Pst I fragment containing the $D_{\beta 2.1}$-$J_{\beta 2.5}$ rearrangement was subcloned into M13mp10 and sequenced by the dideoxy technique utilizing synthetic oligonucleotide primers indicated as open boxes.

(B) Partial restriction map of 3H2.22 and the sequencing strategy used to examine this clone. Vertical boxes and lines above the horizontal line denote exons encoding $D_{\beta 1.1}$, $J_{\beta 2}$, and $C_{\beta 2}$ gene segments. Restriction enzyme abbreviations are as indicated in (A). A 1.3 kb Pst I fragment containing the $D_{\beta 1.1}$-$J_{\beta 2.5}$ rearrangement was subcloned into M13mp10 for sequencing. Synthetic primers utilized in dideoxy sequencing are indicated as open boxes.

Previous studies have demonstrated that incomplete D_β-J_β rearrangements can be expressed as 1 kb transcripts. The $C_{\beta 1}$-specific probe does not detect any 1.0 kb message in T_H 3H.25 RNA. This result is expected because both incomplete D_β-J_β rearrangements involve the $J_{\beta 2}$ gene cluster and hence the $C_{\beta 2}$ gene. When the same filter is hybridized to a $C_{\beta 2}$-specific probe (Kronenberg et

Figure 4. Rearranged Sequences of 3H.25

(A) Sequence of the 3H.25 rearranged variable region gene. The predicted amino acid sequences encoded by the exons are shown above. Differences between 3H.25 and 2B4 (Chien et al., 1984) are listed below. Boundaries between gene segments are indicated by horizontal arrows. The following abbreviations are used: L, leader; V, variable gene segment; D, diversity segment; J, joining segment. N region diversity is underlined.

(B) Sequence of $D_{\beta 1.1}$-$J_{\beta 2.5}$ rearrangement. DNA rearrangement recognition sequences are boxed.

(C) Sequence of $D_{\beta 2.1}$-$J_{\beta 2.5}$ rearrangement.

al., 1985), a low level of 1.0 kb message is detected, which may result from either the $D_{\beta 1.1}$ or the $D_{\beta 2.1}$ rearrangement or both (Figure 5, lane 5). We estimate this transcript is present at less than 10% the level of the 1.3 kb transcript. This low level may reflect differential transcription rates or, alternatively, a lack of stability of the 1.0 kb transcripts. There is only one open reading frame that contains a methionine codon extending 5′ of either the $D_{\beta 1.1}$ or $D_{\beta 2.1}$ gene segments (Siu et al., 1984b). It was suggested that D_β-J_β transcripts could be translated in these frames and produce functional polypeptides. Possible examples of this phenomenon have been reported for B cell heavy chain D-J rearrangements (Reth and Alt, 1984). However, for both D_α-J_α rearrangements in T_H 3H.25, the J_α sequences are not in the same reading frame as these methionine codons. It is unlikely that functional D_α-J_α polypeptides are produced and, accordingly, these transcripts probably play no developmental or functional role in T_H 3H.25.

The T_H 3H.25 cell line appears to have only one functional β-chain rearrangement, an observation that demonstrates the phenomenon of allelic exclusion in the β-chain gene family of this T_H clone. Two other functional T cell lines that we have analyzed by Southern blot analyses appear to have just a single β-chain gene rearrangement each, presumably a functional V_β-D_β-J_β rearrangement (Kronenberg et al., 1985). Therefore, these three T lymphocytes require only one productive β-chain gene rearrangement for the expression of cell surface receptors that recognize antigen and MHC determinants. Further experiments studying individual T cells will be required to determine how general the phenomenon of allelic exclusion is in T cells.

The Same V_β Gene Segment Is Employed by T_H Cells Specific for Different Antigens and Class II Molecules

The rearranged V_β gene segment in T_H 3H.25 cells specific for the antigen HEL and restricted by the class II I-A^b molecule is the same as that of the T_H 2B4 hybridoma, which is specific for the antigen pigeon cytochrome c and restricted by the class II molecule I-E^k (Chien et al., 1984). We have also identified another hybridoma utilizing this V_β gene segment, T_H 11.5, which, like 2B4, is specific for cytochrome c and is restricted by the I-E^k class II molecule. The 11.5 hybridoma also expresses the rearranged 3H.25 V_β gene segment as determined by Northern and Southern blot analyses (Figures 5 and 6). The 3H.25 V_β probe and the $C_{\beta 1}$-specific probe hybridize to a 1.3 kb transcript in T_H 11.5 RNA (Figure 5, lanes 2 and 4), indicating that the transcript containing the T_H 3H.25 V_β gene segment appears to be functional in this cell. The other $J_{\beta 1}$ homolog of the T_H 11.5 cells is joined to $D_{\beta 1.1}$ gene segment in an incomplete rearrangement. Pigeon cytochrome c and HEL are completely distinct polypeptides and the I-A^b and I-E^k molecules differ extensively in their primary structure. Therefore, it appears unlikely that cytochrome c and lysozyme and/or the I-A^b or I-E^k molecules share a common antigenic determinant. This fact was confirmed by demonstrating that T_H 3H.25 did not respond to cytochrome c and T_H 11.5 did not respond to HEL in assays measuring interleukin-2 production in the presence of I-A^b or I-E^k antigen presenting cells (data not shown).

Although the rearranged V_β gene segments in the T_H cells 3H.25 and 2B4 are the same, the variable genes of the β chains are not the same. Different D_β and J_β gene

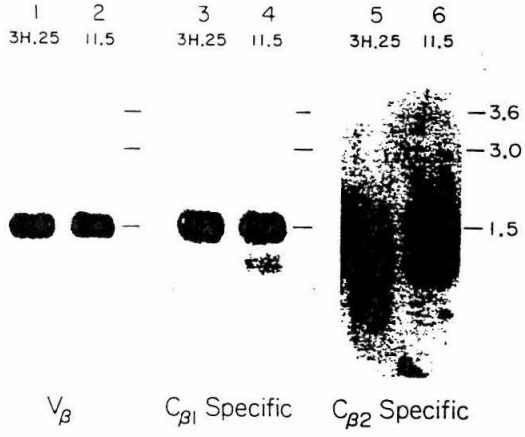


Figure 5. Northern Blot Analysis of Poly(A)$^+$ RNA from T_H 3H.25 and T_H 11.5 Cells

T_H 3H.25 RNA (lanes 1, 3, and 5) and T_H 11.5 RNA (lanes 2, 4, and 6) were hybridized with the indicated probes. The V_β probe is a 250 bp Eco RI–Rsa I fragment purified from 3H1.74. The $C_{\beta 1}$- and $C_{\beta 2}$-specific probes are derived from the 3′ nontranslated sequences of each C_β gene. The $C_{\beta 1}$-specific probe is an Eco RI–Pst I fragment of almost 200 bp isolated from a cDNA clone RBL5-17. The $C_{\beta 2}$-specific probe is a 2.5 kb Sac I fragment purified from the germ-line cosmid clone 2.3W7. The $C_{\beta 2}$-specific probe hybridizes to a 1.3 kb transcript in T_H 11.5 RNA (lane 6), which is derived from expression of the $C_{\beta 2}$ gene in the BW5147 fusion partner in this hybridoma. The migration distances of MS2 phage RNA (3.6 kb) and E. coli ribosomal RNA (3.0 kb and 1.5 kb) are indicated.

segments are used to form each productively rearranged β-chain gene (Figure 4). In addition, these T_H cells may employ different α chains. The specificity for antigen and MHC elements in these two cases must depend not only on the V_β gene segment but also on the distinct D_β and J gene segments used and/or the individual α chains associated with each β chain. This is not a surprising finding as precedent for this observation exists in immunoglobulin genes. A V_H gene of the nitrophenol (NP) family in combination with either a V_λ gene segment or one of two V_κ gene segments is associated with three different antigen specificities—NP, GT (a glutamate-tyrosine polymer) and GA (a glutamate-alanine polymer), respectively (Rocca-Serra et al., 1983). Therefore, it appears likely that the T cell receptor gene families and immunoglobulin gene families rely upon combinatorial manipulation of V gene segments to generate diversity in the repertoire of their receptors.

The 3H.25 V_β Gene Segment Is Used in Other T Cells Specific for Lysozyme and Cytochrome c

The 3H.25 V_β-specific probe was employed in Southern blot analyses of T_H 3H.25 and liver DNAs using three different restriction enzymes (Figure 6A, lanes 1–6). In each case there is one germ-line V_β gene segment in liver DNA and a germ-line and rearranged V_β gene segment in 3H.25 DNA. Thus, in the T_H 3H.25 cells, one 3H.25 V_β gene segment homolog is rearranged and the other is not.

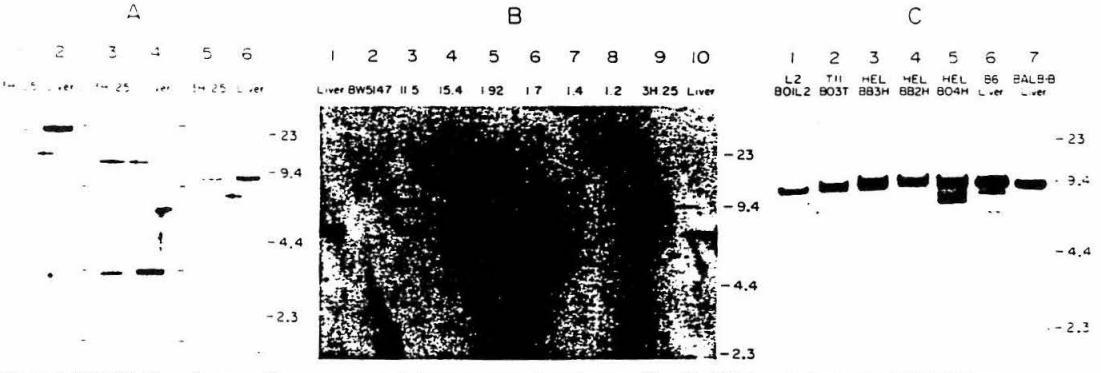


Figure 6. 3H.25 V_β Gene Segment Rearrangements in Lysozyme and Cytochrome c-Specific T Helper Cells and the BW5147 Tumor

(A) T_H 3H.25 and liver DNAs were digested with: lanes 1 and 2, Pvu II; lanes 3 and 4, Bgl II; lanes 5 and 6, Bam HI. Hybridization is with the V_β probe described in Figure 5. Arrows indicate the position of the rearranged band in each lane containing digested T_H 3H.25 DNA. The migration distances of molecular weight markers are indicated.

(B) DNAs from liver, BW5147, T_H cells specific for cytochrome c, T_H cells specific for the T11 peptide of lysozyme, and T_H 3H.25 were digested with Hind III and hybridized with the V_β probe. Lane 1, liver DNA from B10.A5R mice; lane 2, DNA from the tumor BW5147; lanes 3 and 4, DNA from two T_H hybridomas with similar fine specificity for a cytochrome c peptide; lane 5, DNA from a T_H hybridoma specific for acetimidylated-DASp peptide; lanes 6–8, DNA from T_H cells specific for the T11 peptide of lysozyme; lane 9, DNA from T_H 3H.25 and lane 10, liver DNA from C57BL/6 mice. Arrows indicate the position of a rearranged band in T_H 11.5 (lane 3) and T_H 3H.25 (lane 9). The migration distances of molecular weight markers are marked.

(C) DNAs from liver and bulk cell lines specific for HEL or peptides derived from HEL were digested with Bam HI and hybridized with the V_β probe. Lane 1, DNA from a bulk T_H line isolated from C57BL/6 mice immunized with the cyanogen bromide fragment L2; lane 2, DNA from a bulk T_H line obtained from C57BL/6 mice immunized with the tryptic fragment T11; lanes 3 and 4, DNAs from two different bulk T_H lines isolated from BALB.B mice immunized with HEL; lane 5, DNA from a bulk line obtained from C57BL/6 mice immunized with HEL; lane 6, liver DNA from C57BL/6 mice; lane 7, liver DNA from BALB.B mice. The migration distances of molecular weight markers are indicated.

The 3H.25 V_β-specific probe was used in Southern blot analyses on the DNAs from liver, three cytochrome c-specific T_H hybridomas, three T_H clones specific for a tryptic fragment of HEL, T_H 3H.25 cells, and the T cell tumor BW5147 (Figure 6B, lanes 1–10). This probe was also used in Southern blot analyses of the DNAs from six T_C cell lines specific for influenza antigens (data not shown). The 3H.25 V_β gene segment is rearranged only in one of the cytochrome-c-specific hybridomas (T_H 11.5) and the T_H 3H.25 clone. In three T_C influenza lines, one T_H cytochrome c line and two T_H lysozyme lines, the 3H.25 V_β gene segment is found only in the germ-line configuration. Three T_C influenza lines, one T_H cytochrome c hybridoma, the BW5147 tumor, and one T_H lysozyme line appear to have deleted both 3H.25 V_β gene segment homologs. Thus, in six of 14 T cell lines examined, both homologs of the 3H.25 V_β gene segment are deleted. It is possible a few of these T cell lines are haploid for chromosome 6, although this is unlikely since other Southern blots indicate that there are two copies of the $J_{\beta 2}$ gene cluster in at least five of these lines. These observations suggest that two V_β gene segment rearrangements, one on each chromosome, occur relatively frequently in T cells. These results also suggest that the 3H.25 V_β gene segment is 3′ to the V_β gene segments that are rearranged in these six T cells. Thus one may be able to map the relative order of V_β gene segments on chromosome 6 through the use of specific V_β probes and T cell lines that exhibit deletions of both copies of some of the V_β probes.

It is interesting to note that the 3H.25 V_β gene segment is not rearranged in the cytochrome-specific T_H hybridoma 15.4 (Figure 6B, lane 4). This hybridoma was included in our analysis because we wished to determine if the same V_β gene segment would be employed in two different T cells that have the same MHC restriction and fine specificity for antigen. T_H 15.4 is like T_H 11.5 in that it is I-E^k-restricted and has a very similar fine specificity. Both hybridomas recognized pigeon cytochrome c and a 15 amino acid synthetic peptide designated DASp, derived from moth and pigeon cytochrome c sequences (Hansburg et al., 1983). The specificities of the hybridomas can only be distinguished by their reactivities to the DASp peptide containing a glutamine to glutamic acid substitution at a position homologous to residue 100 of cytochrome c. The two hybridomas were both nonreactive or both reactive with peptide antigens containing lysine or serine at position 100. A concordance of activity was also observed when changes at positions homologous to 97, 99, 102, and 103 of the cytochrome c protein were made. Altogether, of 12 different peptides and two proteins tested, T_H 11.5 and T_H 15.4 differed in their reactivity to only one. However, despite the close similarities in antigen specificity and identical MHC restrictions, T_H 11.5 and T_H 15.4 do not employ the same V_α gene. This result suggests that there may be several different T cell receptors that can recognize a single MHC molecule plus an antigen with few determinants such as a small peptide.

We wished to determine if the use of the 3H.25 V_α gene segment could be detected in any other T_H cells whose specificities might be similar to T_H 3H.25. Therefore, we examined the 3H.25 V_β gene segment in bulk T cell lines specific for HEL or various lysozyme peptides. Lysozyme-specific T_H clones derived from inbred C57BL/6 mice can be divided by their response to the lysozyme tryptic peptide T11 (residues 74 to 96)—into groups that recognize T11 and those that do not (N. Shastri and E. Sercarz, unpublished observations). The T_H 3H.25 clone falls into the latter category, whereas T_H clones 1.2, 1.4, and 1.7 (Figure 6B) fall into the former. We immunized inbred BALB.B and C57BL/6 mice with HEL and C57BL/6 mice with lysozyme peptides T11 and L2 (residues 13–105). Bulk T_H cell lines, all restricted to the I-A^b molecule, were established from each of these immunizations. DNA was isolated from these bulk cell lines and used for Southern blot analyses with the 3H.25 V_β probe (Figure 7C). For a specific V_β rearrangement to be detected in a bulk T cell line, either there must be a relatively small number of different T cell clones or many independent clones must employ the same rearrangement. The Southern blot from BALB.B and C57BL/6 germ-line demonstrate the same major 9.4 kb band. The two bulk T_H cell lines derived from BALB.B mice immunized with HEL show no detectable 3H.25 V_β rearrangements (Figure 6C, lanes 3 and 4), whereas the single bulk line from C57BL/6 mice immunized with HEL exhibits a major 3H.25 V_β rearranged band, which migrates as the same size as the rearranged V_β gene in T_H 3H.25 (Figure 6C, lane 5). The size of this band suggests that the 3H.25 V_β gene segment has joined only to $J_{\beta 1}$ gene segments since a different size band would result from a $J_{\beta 2}$ rearrangement with this enzyme. Thus, in spite of the fact that the 3H.25 V_β gene segment is probably identical in BALB.B and C57BL/6 mice (Figure 6C, lanes 6 and 7), in this experiment only the C57BL/6 mice employ any T_H cells in which this V_β gene is rearranged. No 3H.25 V_β rearrangements were detected in bulk T cell lines made from C57BL/6 mice immunized with the T11 and L2 peptides (Figure 7C, lanes 1 and 2). Thus, the 3H.25 V_β gene segment appears to be employed in more than one independently isolated T_H cell in C57BL/6 mice but it is not frequently rearranged in bulk T_H cell lines whose specificities are for different portions of the hen egg white lysozyme molecule. It should be noted that because these lines are probably polyclonal, we do not know the number of different types of T cells represented in each line and must examine cloned T cell lines derived from these bulk lines before we can determine if the 3H.25 V_β gene segment is employed in the major response to HEL in C57BL/6 mice.

The expression of the V_β gene segment repertoire has been addressed in studies in which several V_β gene segments were used as probes to determine clone frequency in a thymus cDNA library (Patten et al., 1984). These V_β gene segments fell into two categories—those expressed frequently and those such as the 2B4 V_β gene segment, which were expressed rarely. However, we have seen the 2B4 V_β gene segment in the two T_H cells, 3H.25 and 11.5, in addition to T_H 2B4, indicating that this V_β gene segment is frequently rearranged and employed in functional T cells with differing specificities. The reason for this discrepancy is not clear. One possibility is that antigen selec-

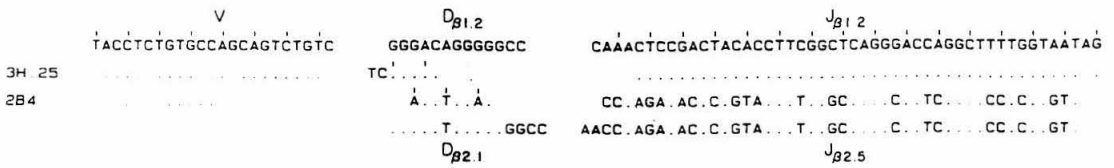


Figure 7. Comparison of the Diversity and Joining Regions of 3H.25 and 2B4

Germ-line sequences of the V, D, and J gene segments used by 3H.25 and 2B4 are shown above and below, respectively, the rearranged sequences. Identities with the upper line are shown with dots. The required reading frame of the V- and J-region sequences is indicated by tics above the first base of each germ-line codon. The different frames used by the D region are similarly indicated above the rearranged sequences.

tion of T cells can significantly perturb V_β gene expression from that found in the thymus. Alternatively, the 2B4 V_β gene segment may have been underrepresented in the thymus cDNA library for technical reasons.

The β-Gene Family Employs a Variety of Mechanisms to Generate Diversity

B cells exhibit three somatic mutational mechanisms—junctional variability arising from the flexibility in the endpoints at which V, D, and J gene segments may be joined, N-region diversity arising from the apparently random trimming and repair of sequences on either side of the D gene segments during joining, and somatic hypermutation, which may occur throughout the rearranged V gene and adjacent flanking regions (Weigert et al., 1980; Kim et al., 1981; Alt and Baltimore, 1982). T cells exhibit at least the first two of these mechanisms, as illustrated by comparison of the three T_H 3H.25 rearrangements to their germ-line counterparts (Figure 7). Both junctional and N-region diversity have occurred in the two incomplete D_β-$J_{\beta 2}$ rearrangements found in T_H 3H.25. The sequence of the $D_{\beta 1}$ gene segment joined to the V_β gene segment is consistent with a V_β-$D_{\beta 1.1}$-$J_{\beta 1.2}$ rearrangement in which seven nucleotides have been deleted from the 3′ end of $D_{\beta 1.1}$ and four nucleotides are deleted from the 5′ boundary of the $J_{\beta 1.2}$ gene segment. Two bases between the joined $D_{\beta 1.1}$ and the V_β gene segments cannot be accounted for in either germ-line gene segment. This indicates either that a $D_{\beta 1}$ gene segment similar but distinct from $D_{\beta 1.1}$ was used or that these two bases were substituted during the rearrangement and represent N-region diversity.

In the two cases studied to date, somatic hypermutation has not been seen in the V_β gene family (Chien et al., 1984; Siu et al., unpublished). The germ-line equivalent of the 3H.25 V_β gene segment has been partially sequenced and is identical through the comparable variable region with both 3H.25 and 2B4 (Figure 4) (Chien et al., 1984). However, the available germ-line sequence does not extend further 5′ into the intron or leader peptide coding exon. Our sequence of the V_β gene segment differs by two bases in the leader exon and two bases in the intron from the published sequence of the rearranged V_β gene segment in 2B4 (see Figure 4). Both exon changes result in amino acid replacements, but do not affect the leader nature of the sequence. Since Southern blot analyses rule out multiple copies of the 3H.25 V_β gene segment (Figure 6A), these differences could be due to polymorphism between strains, somatic hypermutation, or sequencing errors. Polymorphism between B10.A and C57BL/6 mice in the V_β gene segment is not likely because these mice differ at the MHC locus, but have very different genetic backgrounds. Somatic hypermutation is also not likely since these differences are clustered in a region of the V_β gene segment that does not contribute to the β-chain sequence and, accordingly, the T cells expressing this sequence could not be selected for clonal expansion by antigen. Obviously, more extensive analyses will be required to determine whether somatic hypermutation occurs in T cell receptor genes.

Combinatorial joining between mutliple germ-line gene segments contributes fundamentally to immunoglobulin diversity. The same appears true for the β-chain family. The rearrangements of the T_H 3H.25 cell line and other available data suggest that any D_β gene segment may join to any downstream J_β gene segment and that any V_β gene segment may in turn join to any D_β-J_β rearrangement. Analysis of the data that have been published for both functional and incomplete D_β-J_β rearrangements shows that all rearrangements are consistent with either a $D_{\beta 1.1}$ or a $D_{\beta 2.1}$ gene segment joined to a downstream J_β gene segment. While some of these D_β gene segments may be encoded by different germ-line D_β gene segments that have not yet been identified, each of the rearranged D_β sequences could be derived from the characterized $D_{\beta 1.1}$ or $D_{\beta 2.1}$ sequences with the differences accounted for by N-region and junctional diversity. As expected, in each case where the $D_{\beta 2.1}$ gene segment is used (T cells 2B4, CB2, and C5), joining occurs to a $J_{\beta 2}$ gene segment. In T cells in which the $D_{\beta 1.1}$ gene segment is used, half of these rearrangements join the $D_{\beta 1.1}$ gene segment to a $J_{\beta 1}$ gene segment (86T1, 86T5, and 3H.25) and half join to a $J_{\beta 2}$ gene segment (E1, 2C, and 3H.25). Likewise, the 3H.25 V_β gene segment can join to either the $D_{\beta 1.1}$ (3H.25) or $D_{\beta 2.1}$ (2B4) gene segments.

Comparison of the 3H.25 and 2B4 V genes indicates that T cells also generate diversity in a way so far unobserved in B cells. The $D_{\beta 1.1}$ and $D_{\beta 2.1}$ gene segments are similar in sequence and are easily alignable. However, junctional diversity causes the D gene segments to be translated in different translational reading frames in the 3H.25 and 2B4 cells (Figure 7). In fact, among the reported V_β gene sequences, the $D_{\beta 1.1}$ and $D_{\beta 2.1}$ gene segments are translated in all three reading frames. Neither germ-line D_β gene segment codes for stop codons in any of the three translational reading frames. Therefore,

the D_β gene segment may be read in any reading frame as long as the J_β gene segment can be read in its proper frame. In B cells the D_H gene segments exhibit the same types of 5′ and 3′ boundary heterogeneity upon rearrangement. However, for more than 30 V_H gene examples in mice for which multiple uses of a particular D_H gene segment or a closely related homolog could be compared, only one reading frame is employed in spite of the fact that only 3 of the 11 known D_H gene segments have stop codons in at least one reading frame. This restriction on productive D_H rearrangements could reflect selection at the protein sequence level for particular D_H-region residues. Alternatively, the phenomenon could reflect differences in the actual mechanisms employed by T and B cells for gene segment rearrangement.

One final striking observation emerges from these data. Of the seven V_β genes that have been characterized so far, one V_β gene segment has been employed by T_H cells specific for two different antigens (lysozyme and cytochrome c) and different MHC-restricting elements ($I\text{-}A^b$ and $I\text{-}E^k$). Furthermore, most of these V_β gene segments appear to hybridize with only one or two germ-line bands in Southern blot analyses (Patten et al., 1984). The repeated use of the same V_β gene segment in different T cell receptors and the small size of the V_β gene segment families both suggest that the V_β gene repertoire consists of a small number of V_β gene segments whose use is not correlated with recognition of unique antigens or MHC molecules. Therefore, it is not likely that the β chain by itself is responsible for recognition of separate antigen or MHC determinants expressed on the cell surface.

Experimental Procedures

Construction and Screening of the λ Genomic Library

A genomic library was constructed from T_H 3H.25 DNA partially digested with Sau3A1 and the λ bacteriophage vector EMBL3 (Frischauf et al., 1983) according to Maniatis et al. (1982). The library contains about 1×10^6 plaques on 75 filters. Screening of the library was carried out according to Maniatis et al. (1982) using the $J_{\beta 1}$ and $J_{\beta 2}$ probes for hybridization. The filters were hybridized at a probe concentration of 1×10^5 cpm/ml for 18 hr at 65°C in 5× SSC, 5× Denhardt's solution, 10% dextran sulfate, 0.1% SDS, and 100 μg/ml salmon sperm DNA. The filters were then washed in 1× SSC, 0.1% SDS three times for 45 min each at 65°C.

Southern Blots

After restriction enzyme digestion, 10 μg DNA was separated by electrophoresis on a 0.7% agarose gel and transferred to nitrocellulose. Hybridization conditions were the same as those for screening of the library with a probe concentration of 1×10^6 cpm/ml.

Northern Blots

RNA was prepared by resuspension of cells in guanidinium thiocyanate followed by centrifugation through a cushion of CsCl (Chirgwin et al., 1979). Poly(A)$^+$ RNA was selected once on oligo(dT)-cellulose columns and 5 μg was electrophoresed on a 1% agarose formaldehyde gel. The RNA was transferred to nitrocellulose and the filters were prehybridized and hybridized with ^{32}P-labeled nick-translated probes according to Thomas (1980). The filters were washed at 65°C in 1× SSC, 0.1% SDS, three times for 45 min each.

DNA Sequence Analysis

For DNA sequence analyses employing the method of Maxam and Gilbert (1980), 3′-end-labeling was accomplished by filling in a recessed end of a DNA fragment using the large fragment of DNA polymerase I. The 5′-end-labeling was carried out using γ-^{32}ATP and polynucleotide kinase after treatment with calf intestinal phosphatase. Dideoxy sequencing was carried out according to the procedure of Sanger et al. (1977) as modified by E. Strauss (unpublished results) using specific oligonucleotide primers.

Acknowledgments

The authors thank Dr. Suzanna Horvath and Marilyn Tomick for providing synthetic oligonucleotides, and Scheherazade Nasseri for making available lysozyme-specific bulk T cell lines from BALB.B mice, and Mr. Gerald Siu and Drs. Mitchell Kronenberg and Richard Barth for critical reading of the manuscript. We also want to thank Debbie Maloney for technical assistance and Connie Katz, Gwen Anastasi, and Bertha Jones for help in preparing the manuscript. J. G. is supported by a Leukemia Society of America Senior Fellowship. This work was supported in part by T Cell Sciences, Inc.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received December 29, 1984

References

Alt, F. W., and Baltimore, D. (1982). Joining of immunoglobulin heavy chain gene segments: implications from a chromosome with evidence of three D-J_H fusions. Proc. Natl. Acad. Sci. USA *78*, 5812–5816.

Alt, F. W., Rosenberg, N., Enea, V., Siden, E., and Baltimore, D. (1982). Multiple immunoglobulin heavy-chain gene transcripts in Abelson murine leukemia virus-transformed lymphoid cell lines. Mol. Cell. Biol. *2*, 386–400.

Bernard, O., Hozumi, N., and Tonegawa, S. (1978). Sequences of mouse immunoglobulin light chain genes before and after somatic changes. Cell *15*, 1133–1144.

Bernard, O., Gough, N. M., and Adams, J. M. (1981). Plasmacytomas with more than one immunoglobulin κ mRNA: implications for allelic exclusion. Proc. Natl. Acad. Sci. USA *78*, 5812–5816.

Caccia, N., Kronenberg, M., Saxe, D., Haars, R., Bruns, G., Goverman, J., Malissen, M., Willard, H., Yoshikai, Y., Simon, M., Hood, L., and Mak, T. (1984). The T cell receptor β chain genes are located on chromosome 6 in mice and chromosome 7 in humans. Cell *37*, 1091–1099.

Chien, Y., Gascoigne, N., Kavaler, J., Lee, N., and Davis, M. (1984). Somatic recombination in a murine T cell receptor gene. Nature *309*, 322–326.

Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J., and Rutter, W. J. (1979). Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. Biochemistry *18*, 5294–5299.

Clark, S. P., Yoshikai, Y., Taylor, S., Siu, G., Hood, L., and Mak, T. W. (1984). Identification of a diversity segment of human T-cell receptor β-chain, and comparison with the analogous murine element. Nature *311*, 387–389.

Doherty, P. C., and Zinkernagel, R. M. (1975). H-2 compatibility is required for T-cell-mediated lysis of target cells infected with lymphocytic choriomeningitis virus. J. Exp. Med. *141*, 502–507.

Early, P., and Hood, L. (1981a). Mouse immunoglobulin genes. In Genetic Engineering, J. K. Setlow and A. Hollaender, eds. (New York: Plenum), pp. 157–188.

Early, P., and Hood, L. (1981b). Allelic exclusion and nonproductive immunoglobulin gene rearrangements. Cell *24*, 1–3.

Early, P., Huang, H., Davis, M., Calame, K., and Hood, L. (1980). An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: V_H, D and J_H. Cell *19*, 981–992.

Frischauf, A., Lehrach, H., Poustka, A., and Murray, N. (1983). Lambda replacement vectors carrying polylinker sequences. J. Mol. Biol. *170*, 827–842.

Gascoigne, N., Chien, Y., Becker, D., Kavaler, J., and Davis, M. (1984). Genomic organization and sequence of T-cell receptor β-chain constant- and joining-region genes. Nature *310*, 387–391.

Germain, R. N., and Benacerraf, B. (1981). A single major pathway of T-lymphocyte interactions in antigen-specific immune suppression. Scand. J. Immunol. *13*, 1–10.

Golub, E. (1980). Know thyself: autoreactivity in the immune response. Cell *21*, 603–604.

Hansburg, D., Fairwell, T., Schwartz, R. H., and Appella, E. (1983). The T lymphocyte response to cytochrome c. J. Immunol. *131*, 319–324.

Heber-Katz, E., Schwartz, R. H., Matis, L. A., Hannum, C., Fairwell, T., Appella, E., and Hansburg, D. (1982). Contribution of antigen-presenting cell major histocompatibility complex gene products to the specificity of antigen-induced T cell activation. J. Exp. Med. *155*, 1086–1089.

Honjo, T. (1983). Immunoglobulin genes. Ann. Rev. Immunol. *1*, 499–528.

Hunig, T. R., and Bevan, M. J. (1982). Antigen recognition of cloned cytotoxic T lymphocytes follows rules predicted by the altered self hypothesis. J. Exp. Med. *155*, 111–125.

Kappler, J. W., Skidmore, B., White, J., and Marrack, P. (1981). Antigen-inducible, H-2 restricted, interleukin-2 producing T cell hybridomas. J. Exp. Med. *153*, 1198–1214.

Katz, D. H., Hamaoka, T., Dorf, M. E., Maurer, P. H., and Benacerraf, B. (1973). Cell interactions between histoincompatible T and B lymphocytes. J. Exp. Med. *138*, 734–739.

Kavaler, J., Davis, M., and Chien, Y. (1984). Localization of a T-cell receptor diversity-region element. Nature *310*, 421–423.

Kim, S., Davis, M., Sinn, E., Patten, P., and Hood, L. (1981). Antibody diversity: somatic hypermutation of rearranged V_H genes. Cell *27*, 573–581.

Kronenberg, M., Goverman, J., Haars, R., Malissen, M., Kraig, E., Suciu-Foca, N., and Hood, L. (1985). Rearrangement and transcription of the β chain genes of the T cell antigen receptor in different types of murine lymphocytes. Nature, in press.

Kwan, S., Max, E., Seidman, J. G., Leder, P., and Scharff, M. D. (1981). Two kappa immunoglobulin genes are expressed in the myeloma S107. Cell *26*, 57–66.

Lee, N. E., D'Eustachio, P., Pravtcheva, D., Ruddle, F. H., Hedrick, S. M., and Davis, M. (1984). The beta chain of the murine T cell receptor is encoded on chromosome 6. J. Exp. Med. *160*, 905–913.

Malissen, M., Minard, K., Mjolsness, S., Kronenberg, M., Goverman, J., Hunkapiller, T., Prystowsky, M., Yoshikai, Y., Fitch, F., Mak, T., and Hood, L. (1984). Mouse T cell antigen receptor: structure and organization of constant and joining gene segments encoding the β polypeptide. Cell *37*, 1101–1110.

Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982). Molecular cloning, a laboratory manual. (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory).

Maxam, A., and Gilbert, W. (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. Meth. Enzymol. *65*, 499–560.

Patten, P., Yokota, T., Rothbard, J., Chien, K. A., and Davis, M. (1984). Structure, expression and divergence of T-cell receptor β-chain variable regions. Nature *312*, 40–46.

Reth, M. G., and Alt, F. W. (1984). Novel immunoglobulin heavy chains are produced from D-J_H gene segment rearrangements in lymphoid cells. Nature *312*, 418–423.

Rocca-Serra, J., Tonnelle, C., and Fougereau, M. (1983). Two monoclonal antibodies against different antigens using the same V-H germline gene. Nature *304*, 353–355.

Rosenthal, A. S., and Shevach, E. M. (1973). Function of macrophages in antigen recognition by guinea pig T lymphocytes. J. Exp. Med. *138*, 1194–1212.

Sakano, H., Maki, R., Kurosawa, Y., Roeder, W., and Tonegawa, S. (1980). Two types of somatic recombination are necessary for the generation of complete immunoglobulin heavy-chain genes. Nature *286*, 676–683.

Sanger, F., Nicklen, S., and Coulson, A. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA *74*, 5463–5467.

Schwartz, R. H., Yano, A., and Paul, W. E. (1978). Interaction between antigen-presenting cells and primed T lymphocytes: an assessment of Ir gene expression in the antigen-presenting cell. Immunol. Rev. *40*, 153–180.

Siu, G., Clark, S., Yoshikai, Y., Malissen, M., Yanagi, Y., Strauss, E., Mak, T., and Hood, L. (1984a). The human T cell antigen receptor is encoded by variable, diversity, and joining gene segments that rearrange to generate a complete V gene. Cell *37*, 393–401.

Siu, G., Kronenberg, M., Strauss, E., Haars, R., Mak, T., and Hood, L. (1984b). The structure, rearrangement and expression of D_β gene segments of the murine T-cell antigen receptor. Nature *311*, 344–350.

Thomas, D. W., Yamashita, U., and Shevach, E. M. (1977). Nature of the antigenic complex recognized by T lymphocytes. J. Immunol. *119*, 223–226.

Thomas, P. (1980). Hybridization of denatured RNA and small DNA fragments transferred to nitrocellulose. Proc. Natl. Acad. Sci. USA *77*, 5201–5205.

Tonegawa, S. (1983). Somatic generation of antibody diversity. Nature *302*, 575–581.

Weigert, M., Perry, R., Kelley, D., Hunkapiller, T., Schilling, J., and Hood, L. (1980). The joining of V and J gene segments creates antibody diversity. Nature *283*, 497–499.

Reprinted from Nature, Vol. 316, No. 6028, pp. 517-523, 8 August 1985

The murine T-cell receptor uses a limited repertoire of expressed V_β gene segments

Richard K. Barth*, Byung S. Kim*†, Nancy C. Lan*, Tim Hunkapiller*, Nancy Sobieck*, Astar Winoto*, Howard Gershenfeld‡, Craig Okada‡, Dan Hansburg§, Irving L. Weissman‡ & Leroy Hood*

* Division of Biology, California Institute of Technology, Pasadena, California 91125, USA
‡ Department of Pathology, Stanford Medical School, Palo Alto, California 94304, USA
§ Department of Pathology, Fox Chase Cancer Center, Philadelphia, Pennsylvania 19111, USA

Only 10 different V_β gene segments were found when the sequences of 15 variable (V_β) genes of the mouse T-cell receptor were examined. From this analysis we calculate that the total number of expressed V_β gene segments may be 21 or fewer, which makes the expressed germline V_β repertoire much smaller than that of the immunoglobulin heavy-chain or light-chain genes. We suggest that β-chain somatic diversification is concentrated at the V_β-D_β-J_β junctions.

BOTH the T-cell receptor and the immunoglobulins are heterogeneous cell-surface glycoproteins that can recognize many antigens[1-3]. It is clear from genomic analysis that they share similar strategies for diversification. T-cell receptor molecules are composed of α- and β-chains, each of which, like the immunoglobulin chains, is divided into variable (V) and constant (C) regions[1-3]. The V regions together form the antigen-binding domain. The β-chain genes of the mouse are the most thoroughly studied T-cell receptor genes. Like the immunoglobulin genes, they are divided into separate V_β, diversity (D_β) and joining (J_β) gene segments that are assembled by recombination during T-cell development to form a V_β gene that is associated with either of two constant ($C_\beta 1$ and $C_\beta 2$) genes[4-8]. There are six functional J_β gene segments clustered just upstream of each C_β gene[4-6] and two D_β gene segments, $D_\beta 1.1$, upstream of the $J_\beta 1$ cluster, and $D_\beta 2.1$, upstream of the $J_\beta 2$ cluster[7,8]. The total number of V_β gene segments is unknown. Like immunoglobulins, the T-cell receptor β-chain has at its disposal three sources of diversity: a multiplicity of germline gene segments[4-9]; combinational diversity through the assembly of different V, D and J segments[4,9-11]; and somatic mutation[9,11]. Immunoglobulin genes have three sources of somatic mutation: junctional flexibility at the sites of gene-segment joining[12-14]; the addition of random nucleotides to either side of the D-gene segment in the process of joining (N-region diversity)[15]; and somatic hypermutation[16,17]. It is known that β-chain genes use the first two processes but may not use the third[4,7-9,11].

T-cell antigen recognition differs from that mediated by immunoglobulins in that T cells must recognize antigen in the context of a cell-surface molecule encoded by the major histocompatibility complex (MHC), a phenomenon termed MHC restriction[18,19]. T-cytotoxic (T_C) cells, which are capable of killing virus-infected and tumour cells, are mainly restricted by class I gene products of the MHC[20]. T-helper (T_H) cells, which are capable of enhancing B- or T-cell responses, are mainly restricted by class II MHC gene products[21]. T-cell receptor diversification must therefore accommodate antigen recognition and recognition of highly polymorphic determinants on MHC molecules.

To determine the extent of T-cell receptor diversity and its relationship to antigen/MHC recognition, we analysed eight V_β genes from complementary DNA libraries of functional T cells and thymocytes. We have compared these V_β gene sequences with seven from the literature and find that: (1) the expressed V_β gene repertoire is probably small, perhaps less than 21 members; (2) V_β protein segments are structurally similar to immunoglobulin V segments; and (3) there is no simple correlation between antigen and MHC specificity and the use of particular β-chain gene segments.

Expression of V_β gene segments

We determined the nucleotide sequence of eight V_β genes obtained from cDNA libraries that were constructed from thymus cells, the T_H hybridoma 1.9.2 specific for the cytochrome

† Permanent address: Department of Microbiology/Immunology, Northwestern University, Chicago, Illinois 60611, USA.

Table 1 Characteristics and origins of sequenced V_β genes

V_β gene	Class	Strain	Antigen/MHC specificity	V_β	D_β	J_β	Ref.
2B4	T_H	B10.A	Cytochrome c/I-$E^k_\alpha E^k_\beta$	3	2.1	2.5	4
1.9.2	T_H	B10.A[5R]	AmDASp/I-E^k (see ref. 22)	1	1.1	1.1	*
3H.25	T_H	C57BL/6	Hen-egg lysozyme/I-A^b	3	1.1	1.2	11
C5	T_H	C57BL/6	Dinitrophenol-ovalbumin/I-A^b	8.1†	2.1	2.5	9
E1	T_H	BALB/c	Trinitrophenol/I-A^d	2	1.1	2.2	9
LB2	T_H	C57BL/6	Chicken red blood cell/I-A^b	6	2.1	2.3	9
HDS11	T_C	BALB.B	H-2^d	7	1.1	2.6	24
AR1	T_C	C57L	H-2^d	2	1.1	2.5	*
86T1	Thymocyte	BALB/c	—	1	1.1	1.3	23
TB2	Thymocyte	C57BL/Ka	—	8.2†	2.1	2.5	*
TB3	Thymocyte	C57BL/Ka	—	4	2.1	2.5	*
TB12	Thymocyte	C57BL/Ka	—	8.1	1.1 or 2.1‡	2.4	*
TB21	Thymocyte	BALB/c	—	5.1	2.1	2.6	*
TB23	Thymocyte	BALB/c	—	8.3†	ND	ND	*
BW5147	Tumour	AKR	—	1	2.1	2.5	*

* This paper.
† The three members of the $V_\beta 8$-subfamily are denoted $V_\beta 8.1$, $V_\beta 8.2$ and $V_\beta 8.3$.
‡ So little of the D_β gene segment remains in the rearranged V_β gene that it is impossible to know which D_β gene segment contributed the sequence. ND, not determined.

c peptide AmDASp (ref. 22) and a functional T_C-cell line specific for MHC alloantigens. The sequence of these V_β genes and seven additional V_β genes taken from the literature were then compared (Fig. 1*a*, Table 1)[4,9,11,23,24]. Of the 15 V_β genes analysed, 6 are from T_H cells, 2 from T_C cells, 6 from thymocytes and 1 from a T-cell tumour (Table 1). The antigen/MHC specificities of the functional cells are listed in Table 1. The specific functions and antigen/MHC specificities of the T-cell tumour and the cells that generated the thymocyte-derived V_β genes are unknown. As 99% of thymocytes do not migrate out of the thymus[25], some of the thymocyte-derived V_β genes may come from non-functional T cells. The 10 distinct protein sequences of the 15 V_β gene segments are shown in Fig. 1*b*: there are 10 distinct sequences. For reasons discussed below, the first seven sequences listed in Fig. 1*b* are designated V_β1–7 and the last three V_β8.1, V_β8.2 and V_β8.3. One V_β gene segment appears three times (V_β1), three appear twice (V_β2, V_β3, and V_β8.1) and six appear once (V_β4, V_β5.1, V_β6, V_β7, V_β8.2 and V_β8.3). Moreover, when the sequences of seven additional V_β genes isolated from six T_H cells by our laboratory[26] and others are included (J. Goverman, unpublished data; M. Steinmetz, personal communication; A. Fotedar, P. Morinaga, B. Singh, T. Tamaoki and T. Wegmann, personal communication; S. Hedrick, personal communication), 11 different V_β gene segments are found in 22 rearranged genes. Six of the eleven different V_β gene segments appear more than once. Common V_β gene segments occur in functional T cells, thymocytes and the T-cell tumour BW5147. The repeated use of V_β gene segments in a sample of this size suggests that the repertoire of expressed V_β gene segments is small. If we assume that the V_β gene segments are expressed randomly, the size of the V_β gene segment repertoire can be determined statistically.

If there are L species of V_β gene segments and they are expressed with equal probability and are randomly selected

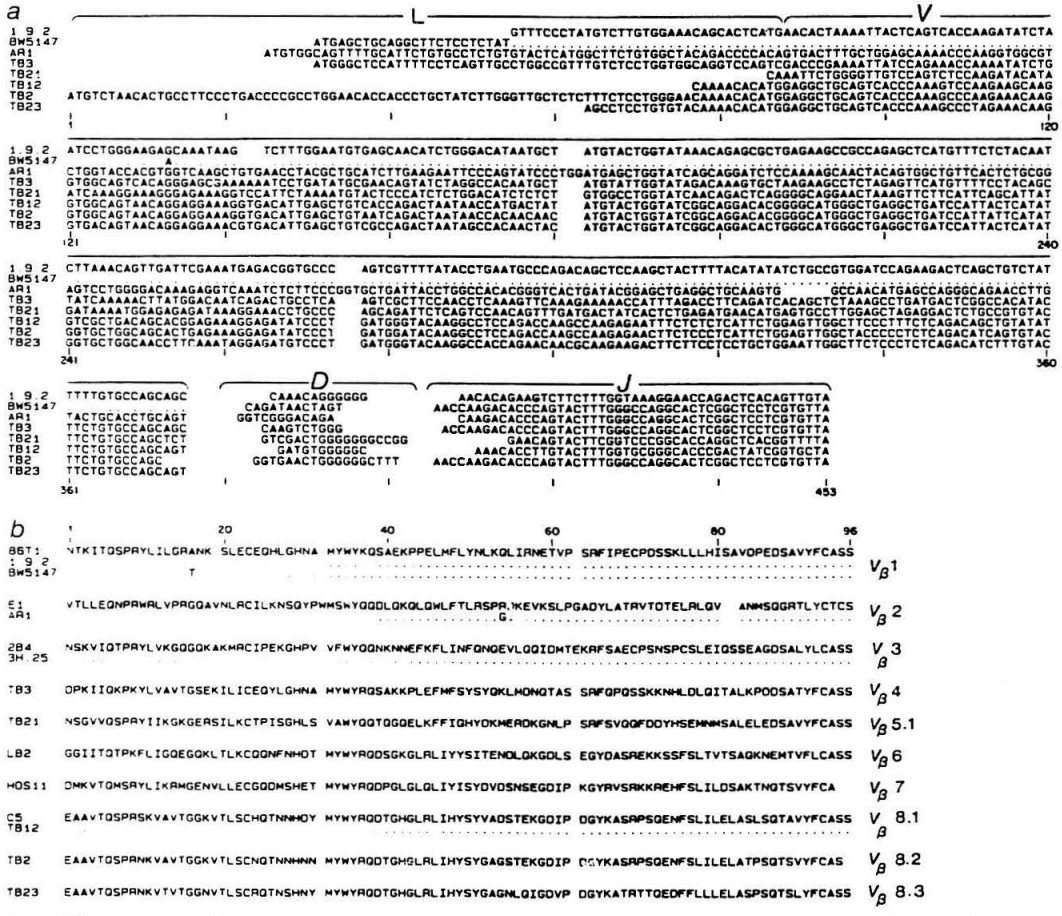


Fig. 1 *a*, DNA sequence of eight V_β genes. All cDNA libraries were constructed in the λgt10 cloning vector by a modification of the method of Huynh *et al.*[56]. Libraries were screened with a ^{32}P-labelled C_β cDNA probe, positive clones were isolated and cDNA inserts were subcloned into the M13 mp8 or mp10 vectors. The V_β genes of each subclone were sequenced on both strands by a modification of the dideoxynucleotide-chain terminating method of Sanger[57], using specific oligonucleotide primers[58]. The sequences are displayed with the V_β and J_β gene segments and *D* regions aligned separately. L, nucleotides encoding the leader peptide. The TB23 subclones contained only a V_β gene segment. Dots, identity between the V_β segments used by 1.9.2 and BW5147. The cDNA obtained from AR1 contained only a portion of a V_β gene segment beginning at position 161. For purposes of sequence alignment, the 5′ portion of the homologous V_β gene segment published previously[9] has been added. *b*, Translated protein sequences of 15 V_β gene segments. The DNA sequences of the eight V_β gene segments shown in *a*, as well as seven V_β gene segments published previously, were translated and aligned to each other to maximize sequence homology. (See Table 1 for the source of each sequence.)

from the pool, then the probability (P) of obtaining an observed result is

$$P(m_1, m_2, \ldots, m_N) = \frac{1}{L^N} \frac{L!}{(L-M)!M!} \times \frac{N!}{(1!)^{m_1}(2!)^{m_2}\ldots(N!)^{m_N}} \times \frac{M!}{m_1!m_2!\cdots m_N!}$$

where m_1 is the number of different species found once; m_2, the number of different species found twice; m_N, the number of different species found N times; $N = \sum_{i=1}^{N} im$ = the total number of samples analysed; and $M = \sum_{i=1}^{N} m_i$ = the number of different species found in the sampling. Conversely, given the observed result, the relative likelihood that there are exactly L species in the pool is

$$P(L) = \frac{1}{\sum_{L'=M}^{\infty} \frac{L'!}{L'^N(L'-M)!}} \frac{L!}{L^N(L-M)!}$$

For the data discussed above, $N = 22$ and $M = 11$. By choosing a range of values for L and testing each for the probability of arriving at the value $M = 11$ for $N = 22$, the size of the expressed V_β gene segment family is 21 or less at the 95% confidence level. This estimate of the mouse V_β-gene segment repertoire is much smaller than the mouse immunoglobulin V_H (~100–300) or V_κ (~100–300) gene-segment families[27,28], but larger than the mouse V_λ (2) repertoire[29]. However, mouse λ-chains are expressed in only a few per cent of mature B cells, whereas β-chains are expressed in all T_H and T_C cells that have been analysed[30,31].

We do not know whether each V_β gene segment is expressed with equal probability in the T-cell population or if the sampling of V_β genes is random. In fact, the relative occurrence of V_β gene segments that we observe can also be explained by the frequent use of a small subset of V_β gene segments. Therefore, a much larger set of V_β gene segments could be expressed infrequently. However, the fact that we find identical V_β gene segments expressed in T cells that differ in their antigen recognition and MHC restriction as well as between functional T cells and unselected thymus cells (Table 1), indicates that the effective repertoire of expressed V_β gene segments is probably very small. Accordingly, a large multiplicity of germline V_β gene segments does not seem to be a major contributing factor in the generation of T-cell receptor diversity.

Single gene subfamilies

Another method of estimating the size of the V_β gene segment repertoire is to determine the number of V_β gene segments in the mouse genome that cross-hybridize with V_β probes. This technique will identify V_β gene segments that have extensive homology with the available V_β gene segment probes, and gives a minimum estimate that can be compared with the predicted number. When this type of analysis was carried out with mouse V_H and V_κ gene segment probes, all the V gene segments fell into one of several distinct multigene subfamilies[27,28,32]. In mice there seem to be at least 7 V_H subfamilies ranging in size from 2 to >40 members and at least 5 V_κ subfamilies ranging in size from 2 to >20 members[27,28,32]. (V gene segments with ≥75% similarity have been defined as belonging to the same subfamily[33].) From this analysis it was estimated that the total repertoire of V_H and V_κ gene segments is ~100–300 members for each family[27,28].

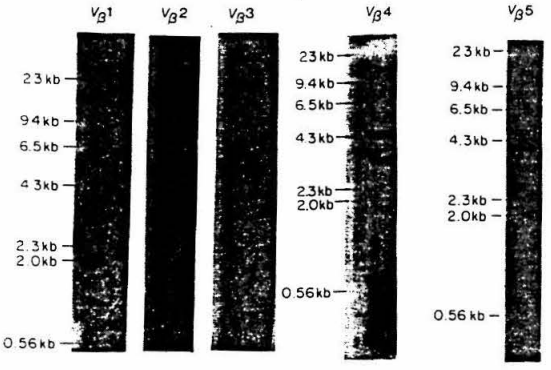


Fig. 2 Southern blot analysis of mouse germline DNA using probes from five different V_β gene segment subfamilies. BALB/c mouse liver DNA was digested with *Eco*RI (V_β1, V_β4, V_β5 and V_β8) or *Bam*HI (V_β2). 10 μg of each DNA digest was electrophoresed on a 0.8% agarose gel and transferred to nitrocellulose. Blots were hybridized with one of five ^{32}P-labelled V_β gene segment probes at 65 °C in 1 M NaCl, 40 mM Tris, *p*H 7.5, 10% dextran sulphate, 1 × Denhardt's solution, 0.1% SDS and 100 μg ml^{-1} denatured salmon-sperm DNA. The blots were then washed at 65 °C in 0.15 M NaCl, 0.015 M sodium citrate, *p*H 7.5, and 0.1% SDS. The probes are derived from subclones of each V_β gene segment. The faint 5.7-kilobase (kb) band in the V_β2 blot results from contamination of the probe with J_β region sequences.

Table 2 shows the percentage similarity between the 10 different V_β gene segments at the protein and nucleotide level. The nucleotide similarity of the V_β gene segments used by the TB2, TB12 and TB23 V_β genes range from 85 to 92%, indicating they are members of the same V_β subfamily. These V_β gene segments have been designated V_β8.1, V_β8.2 and V_β8.3, respectively (Fig. 1*b*). The similarities among the remaining V_β gene segments range from 34 to 63%. Therefore, these V_β gene segments belong to different V_β subfamilies, which we have denoted V_β1–V_β7. As the expressed V_β gene segment repertoire appears limited, it is important to have a generally accepted nomenclature for these sequences, as used with V_L and V_H subgroups[34].

Table 2 Homology matrix of the 10 V_β gene segments

	V_β8.1	V_β8.2	V_β8.3	V_β6	V_β7	V_β1	V_β3	V_β4	V_β2	V_β5
Subfamily										
V_β8.1	—	90	77	47	52	29	26	28	27	29
V_β8.2	92	—	81	45	52	29	24	28	26	26
V_β8.3	85	88	—	41	48	28	23	28	29	24
V_β6	55	54	54	—	43	27	27	29	25	22
V_β7	61	63	60	53	—	28	23	27	23	29
V_β1	45	43	44	45	49	—	33	53	20	37
V_β3	46	47	47	46	48	50	—	30	18	37
V_β4	46	46	46	47	47	62	50	—	20	33
V_β2	42	44	42	39	39	34	38	38	—	18
V_β5	45	46	43	41	48	53	54	47	38	—

Numbers above the diagonal designate the percentage similarity of sequences on the x and y axes when compared at the protein level; numbers below the diagonal show percentage similarity at the DNA level.

To determine the size of the different V_β subfamilies, Southern blot analysis was performed on mouse liver DNA using DNA probes for the $V_\beta 1$, $V_\beta 2$, $V_\beta 4$, $V_\beta 5$ and $V_\beta 8$ subfamilies (Fig. 2). Three of the V_β probes show a single band, indicating that each V_β gene segment represents a different single-gene segment subfamily ($V_\beta 1$, $V_\beta 2$ and $V_\beta 4$). The $V_\beta 5$ and $V_\beta 8$ subfamilies appear to have two and three members, respectively (Fig. 2). Because the V_β gene segment used by the TB21 V_β gene is the first isolated member of the $V_\beta 5$ subfamily, we have denoted it $V_\beta 5.1$ (Fig. 1*b*). It has been previously reported that the $V_\beta 1$, $V_\beta 2$, $V_\beta 3$, $V_\beta 6$ and $V_\beta 7$ subfamilies are single genes and that the $V_\beta 8$ subfamily has two members[9,11,24]. Thus, six different V_β subfamilies with one member, one V_β gene segment subfamily with two members and one V_β subfamily with three members have been identified. Including the additional V_β gene segment characterized by Malissen *et al.*[26], there are at least 12 mouse V_β gene segments. This minimum estimate for the size of the V_β gene segment family falls within the range indicated by the statistical analysis presented above and is consistent with the hypothesis that the V_β gene segment repertoire is small.

The V_β gene segment family with its six single-member subfamilies differs from those of the immunoglobulin V_H and V_κ gene families, each of which contains 5 or more subfamilies with 2–40 or more members[27,28,32]. Southern blots of DNAs from several rodent species as well as rabbit and human analysed with various V_β probes indicate that the single-member V_β subfamily sequences are less conserved between species than those of the three-member family ($V_\beta 8$)[9]. It has been suggested that this difference reflects selective pressures on the single-copy sequences to diverge rapidly, presumably to accommodate recognition of antigen in a changing MHC context[9]. An *ad hoc* argument for specific, positive selection must then be made to explain the unique conservation of the $V_\beta 8$ subfamily. In apparent conflict with this view is the observation that very little restriction enzyme polymorphism of single copy V_β sequences is seen between mouse strains that have diverged significantly in their MHC genes (B.S.K. and R.K.B., unpublished observations; D. Loh, personal communication). In addition, the inbred SJL mouse has deleted two of the five V_β gene segment subfamilies examined (B.S.K. and R.K.B., unpublished observations; D. Loh, personal communication). This indicates that mice have distinct V_β haplotypes containing different combinations of V_β subfamilies. Hence, an alternative explanation for the lack of V_β-gene segment conservation between species is that the ancestor to mammals contained different V_β haplotypes and during speciation distinct V_β haplotypes were passed on to different evolutionary lines. This model does not require a high rate of V_β gene segment mutation and consistent with the lack of restriction enzyme V_β polymorphism in mice. Additional data on the evolutionary divergence of the V_β gene segments should clarify the explanations for these observations.

It is unclear why single-member subfamilies seem to have arisen exclusively in the V_β gene family. Note that no V_β pseudogenes have been found, whereas at least 30% of V_H gene segments are pseudogenes[27]. These observations raise a question about the mechanisms that are responsible for retarding the duplicative processes seen in other V gene families.

Diversification mechanisms

Despite the fact that a limited number of V_β gene segments have been identified, all 15 of the V_β gene sequences examined are distinct from one another because of combinatorial and somatic mutational processes (Fig. 1*a*)[4,7–11]. The germline, combinatorial and somatic mutation contributions to V_β gene assembly are summarized below.

Germline. The β-chain gene family differs from its V_H counterpart in the apparently limited number of germline V_β and D_β gene segments, the expressed V_β gene segment repertoire consisting of perhaps 21 or fewer members compared with 100–300 germline V_H gene segments. In addition, although we do not know the number of germline D_β gene segments, comparisons

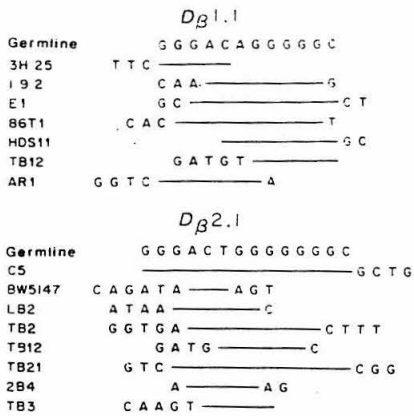


Fig. 3 The D regions from 15 V_β genes may arise from just two D_β gene segments. The D regions shown in Fig. 1*a* and those published previously[4,9,11,23,24] were aligned to either $D_\beta 1.1$ or $D_\beta 2.1$. The regions homologous to either D_β gene segment are represented by a straight line. The additional nucleotides flanking the germline D_β sequences are presumed to be added by N-region diversity. The D region used by TB12 can be derived from either $D_\beta 1.1$ or $D_\beta 2.1$ and is so indicated.

of the D segments found in rearranged V_β genes indicates that all could be derived from the two D_β gene segments previously identified[7,8] assuming extensive junctional flexibility and N-region diversification (Fig. 3). The heavy-chain locus, by contrast, has at least 10–20 D_H segments[35]. Finally, there are 12 apparently functional J_β gene segments, 6 in each J_β gene cluster[4–6], compared with 4 J_H gene segments[36–38].

Combinatorial. Combinatorial joining permits either D_β gene segment to be joined to any downstream J_β gene segment ($6\ D_\beta 1 J_\beta 1 + 6 D_\beta 1 J_\beta 2 + 6 D_\beta 2 J_\beta 2 = 18 D_\beta - J_\beta$ rearrangements). Individual V_β gene segments appear to join any D_β–J_β rearrangements ($21 \times 18 = 378$ V_β genes). The $D_\beta 1$ and $D_\beta 2$ sequences are each used approximately half the time in the sample analysed and $D_\beta 1$–$J_\beta 2$ joinings occur as frequently as $D_\beta 1$–$J_\beta 1$ joinings (3 against 3)(Table 1). Eight different $J_\beta 1$ and $J_\beta 2$ gene segments are used, with $J_\beta 2$ gene segments being used in 11 out of 14 examples. Thus, one would expect individual V_β gene segments to join with the $J_\beta 1$ gene cluster 25% of the time and the $J_\beta 2$ gene cluster 75% of the time, which is what is found, supporting the contention that the joining of the $D_\beta 1.1$ gene segment to either J_β cluster occurs randomly. However, the $J_\beta 2.5$ gene segment is used 6 of the 11 times that the $J_\beta 2$ cluster is used. Whether this bias represents the selective effects of antigen or the mechanisms involved in DNA rearrangement is uncertain.

There may be two additional combinatorial mechanisms. Asymmetrical recognition sequences surrounding the D_β gene segments potentially permit V_β–J_β and D_β–D_β joinings[7,8]. Data consistent with the former possibility have been presented[39], although the interpretation of these data is difficult because of the possible loss of D gene sequences during V_β gene segment rearrangement. As yet there is no evidence for either mechanism; hence, if these joinings do occur, they are infrequent.

Somatic mutation. Junctional flexibility in joining gene segments is illustrated in Figs 1*a* and 3. One to six extra nucleotides are found at either end of the D gene segments. Interestingly, there appears to be no G/C bias (50%) in this N-region diversification in contrast to that reported for immunoglobulin N-region diversity[15].

The D_β gene segments may join to the V_β gene segments with equal probability in all three translational reading frames, with

the only requirement being that the D_β-J_β and V_β-D_β joinings leave the J_β sequence in the proper translational frame (data not shown)[11]. In contrast, examination of >75 productively rearranged V_H genes indicate that D_H-gene segment families show a strong preference for joining in one translational reading frame (85%)(M. Kaartinen and O. Mäkelä, personal communication; T.H., unpublished observations). Presumably this difference arises as a consequence of more rigid structural or selective constraints on the D_H as opposed to D_β segments.

The somatic hypermutation of immunoglobulin genes occurs late in B-cell development, perhaps on exposure to antigen[40,41]. By contrast, two V_β gene segments expressed in functional T_H cells specific for different antigens are identical to the germline sequence[4,11]. On the other hand somatic variants can arise in alloreactive T cells *in vitro*[42], although the physiological relevance of this observation is uncertain. We see two nucleotide substitutions and one replacement among the three V_β1, and two nucleotide substitutions and one replacement between the two V_β2 gene segments that have been sequenced (Fig. 1*a*, *b*)[9,23]. As all these V_β sequences are derived from different strains of mice (Table 1), these differences may result from polymorphism. No other differences were found and we tentatively conclude, in agreement with earlier workers[4,11], that somatic hypermutation, if it exists at all, is much less extensive in V_β than in immunoglobulin *V* genes.

V-region similarities

It has been suggested that both the total diversity and the distribution of variability of V_β segments differ from those of immunoglobulin V segments and that this may reflect the additional requirements for MHC recognition[9]. We have analysed the pattern of sequence diversity of the available V_β sequences with this in mind.

The percentage similarities at the protein level (Table 2) between the different V_β subfamilies ranges from 18 to 53% and 77 to 90% between members of the V_β8 subfamily. It has been suggested that V_β segments are substantially more divergent than V_H sequences[9]. Although their range extends to a lower value (18%) than that observed so far for known mouse V_H segments (34%), or human V_H segments (24%)(data not shown), when sampling biases are considered (see below) the maximum variation between different V_β segments (82%) does not seem significantly larger than the maximum variation found between immunoglobulin V subfamilies (76%).

In immunoglobulin V regions, most of the sequence variability is clustered within three specific 'hypervariable' regions that form the antigen-binding crevice of antibody molecules[43–46]. Two of these hypervariable regions are encoded by the *V* gene segments and the third is found in the V_H-D_H-J_H or V_L-J_L junction regions[47]. We have compared the pattern of variability of the 10 translated mouse V_β gene segments (Fig. 1*b*) with that of a set representing all the known human V_H segments and a set of 18 human V_H segments with blocked α-amino groups (Fig. 4). Human V_H segments were chosen for comparison because they offer a more random representation of V_H sequences than any set of mouse *V* segments that has been sequenced. This is because mouse V_H sequences have been highly selected in comparison with V_β sequences in two ways. First, mouse V_H sequences are derived mainly from immunoglobulins that recognize a relatively limited number of antigens. Second, for technical reasons, mouse V_H sequences are almost exclusively determined from heavy chains with unblocked α-amino groups despite the fact that 80% of mouse serum immunoglobulins have heavy chains with blocked α-amino groups[48]. The blocked human V_H sequences, on the other hand, were selected randomly from various tumours and patients with other pathological conditions[47]. We have found that the variability distribution of the mouse V_β segments (Fig. 4*a*) is very similar to the distribution found for the set of human α-amino blocked V_H segments (Fig. 4*b*). The variability distribution of the total set of human V_H segments (Fig. 4*c*) represents a less random sampling than the human α-amino-blocked V_H segments for the same reasons

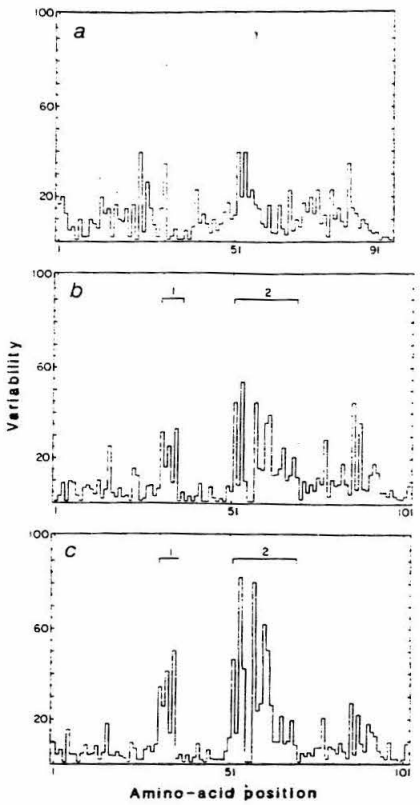


Fig. 4 Variability plot of V_β and V_H segments. Variability at each amino-acid position *N* is calculated[43] as: variability$_N$ = no. of different amino acids that occur at *N*/frequency of most commonly occurring amino acid at *N*. Hypervariable regions can be defined empirically as a set of residue positions whose average variability is substantially greater than the mean variability of the entire sequence. The sequences used are all those available from the protein information resource of the National Biomedical Resource Foundation and GenBank. *a*, Translated sequence of the 10 distinct V_β gene segments shown in Fig. 1*b*; *b*, 18 α-amino-blocked human V_H sequences; *c*, 31 human V_H sequences representing sequences blocked and unblocked at the α-amino position.

given for the mouse sequences, and shows a more accentuated variability at the two classically defined hypervariable regions. The total set of human V_H sequences also exhibits a slightly lower background variability, partly resulting from the larger size of the sampling. Thus, it is critical to consider sample size and selection when this type of analysis is conducted. We conclude that the distributions of variability in the V_β and V_H segments are not significantly different from one another. Our results are not consistent with the suggestion[9] that V_β regions have novel hypervariable regions relative to immunoglobulin V regions.

We have also compared V_β and immunoglobulin V segments by analysing them for two properties believed to reflect important structural features of these molecules, the distribution of β-pleated sheet-forming potential[49] and the predicted hydrophobicity profile[50]. We find the results of these analyses to be almost identical for mouse V_β, V_H and V_κ segments (Fig. 5). Patten *et al.* have also reported that the β-pleated sheet patterns of several V_β segments conform to that of representative V_H and V_κ sequences[9]. V_β sequences also conserve essentially the

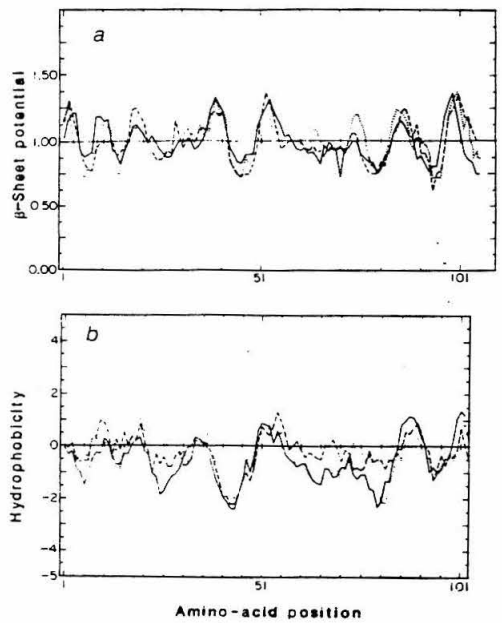


Fig. 5 Secondary structure analyses of V_β, V_H and V_α segments. *a*, β-pleated sheet potential plots, using the method of Chou and Fassman[49]; *b*, hydrophobicity plots using the scale of Kyte and Doolittle[50]. Solid lines, β; dotted lines, H; dashed lines, κ. Analyses in *a* and *b* were based on the average value at each position for 55 V_H regions, 100 V_α regions and 10 V_β regions.

same group of residues that in immunoglobulins are thought to be important in intra-chain structural interactions (data not shown)[51,52]. These results strongly support the contention that the general biochemical characteristics and predicted secondary structures of the V_β, V_H and V_α regions are very similar to one another. Accordingly, we predict that the T-cell receptor and immunoglobulin molecules fold into comparable tertiary structures. Therefore, from the analysis of the sequence variability and structural predictions, we find no evidence for the existence of any fundamental differences in how V_β gene segments can contribute to determinant recognition when compared with immunoglobulin V gene segments. This conclusion is supported by the observation that MHC-restricted antibodies have been raised to influenza antigens[53]. This observation implies that there need be nothing structurally unique about the T-cell receptor structure and its ability to recognize antigen and MHC-restricting elements.

Antigen/MHC specificity

T cells can recognize a range of antigens similar to those recognized by B cells, but in conjunction with the entire range of polymorphic MHC molecules found in a species. Therefore, the T-cell receptor repertoire is expected to be at least equal to the immunoglobulin repertoire. Our observations suggest that the β-chain genes use fewer V gene segments than do immunoglobulins and that somatic hypermutation is infrequent or even non-existent. On the other hand, the V_β subfamilies exhibit the same overall range of sequence diversity as different V_H subfamilies. Furthermore, the contribution of J_β gene segments is three times that of J_H or J_α gene segments, and because the D_β gene segments can be used readily in all three translational reading frames, the smaller family size than is seen in immunoglobulins is at least partially compensated for. Finally, β-chains and immunoglobulins both use junctional flexibility and *N*-region diversification, which are major contributors to the generation of diversity within a specific region of the molecule involved in determinant recognition. Hence, we conclude that the limited number of expressed V_β gene segments and the infrequent use of somatic hypermutation does not necessarily reflect a restricted β-chain repertoire. Rather, these observations imply that β-chain somatic diversification is more focused to the 3′ portion of the V_β gene.

There are three possible explanations for the apparent low frequency of somatic hypermutation of V_β genes. First, somatic hypermutation may act in B cells to increase the antigen binding affinity of antibodies rather than to generate a broader range of antigen response[41,54]. T-cell receptors may not need somatic hypermutation because they may operate with a lower binding affinity due either to a lower affinity requirement for T-cell response or to the stabilizing effect of accessory molecules on the T-cell surface. Second, the current sampling may not include T cells equivalent to secondary-response B cells. Third, somatic hypermutation may have highly unfavourable consequences in T cells. In B cells somatic hypermutation occurs late in development, after antigen stimulation. It is possible that somatic hypermutation does not occur late in T-cell development, after immunocompetent cells have migrated from the thymus, in order to prevent the generation of autoreactive regulatory or cytotoxic T cells in the periphery. Thus, somatic diversification may be restricted to occur early in T-cell development, which allows the thymus to remove autoreactive cells arising from the somatic variation. B cells may not be so restricted because of the strong influence of regulatory T cells[55], and thus are free to undergo somatic hypermutation later in B-lymphocyte development in response to antigen stimulation.

As Table 1 demonstrates, there is no simple correlation between a particular V_β gene segment and distinct antigen specificities or MHC-restricting elements. For example, the $V_\beta 2$ gene segment is used by a T_H-cell specific for trinitrophenol and the I-A^d MHC molecule (E1) and a T_C specific for the H-$2D^d$ alloantigen (AR1). If the V_β and V_α regions fold in a manner similar to that of their immunoglobulin counterparts, as is suggested for V_β regions by our earlier analysis, then both chains will play a critical role in generating the binding site for antigen plus MHC. Accordingly, there is no reason to believe that either chain will have a particular role in recognizing either antigen or MHC individually.

We thank Marcus Meister for deriving the statistical calculation used for V_β gene segment number estimates and help with statistical analyses. We also thank Susan Stone, Debbie Maloney and Marty Garcia for technical assistance, Gerald Siu and Dr Mitchell Kronenberg for helpful advice and stimulating discussion, Drs Nilabh Shastri, Joan Kobori, Nancy Costlow, Joan Goverman, Stephen Hunt, Regina Haars, Ulf Landegren and Martha Zuniga for critical comments, and Gwen Anastasi, Connie Katz and Bertha Jones for help in preparing the manuscript. We acknowledge the following sources of support: the Cancer Research Institute (R.K.B.) and the NIH (L.E.H.).

Note added in proof: Recently, the sequences of 15 additional V_β genes have been determined (D. Loh, personal communication). When these are added to the 22 existing V_β sequences, only 16 different V_β gene segments are defined from a total of 37 sequences. These additional data support our hypothesis of a limited V_β gene segment repertoire. Furthermore, when these data are included with our own data and analysed for the distribution of variability, hydrophobicity and β-pleated sheet-forming potential, the results are very similar to those presented here.

Received 10 April; accepted 13 June 1985.

1. Allison, J. P., McIntyre, B. & Bloch, D. *J. Immun.* **129**, 2293–2300 (1982).
2. Haskins, K. *et al. J. exp. Med.* **157**, 1149–1169 (1983).
3. Meuer, S. C. *et al. J. exp. Med.* **157**, 705–719 (1983).

4. Chien, Y.-H., Gascoigne, N. R. J., Kavaler, J., Lee, N. E. & Davis, M. M. *Nature* **309**, 322-326 (1984).
5. Malissen, M. *et al. Cell* **37**, 1101-1110 (1984).
6. Gascoigne, N. R. J., Chien, Y.-H., Becker, D. M., Kavaler, J. & Davis, M. M. *Nature* **310**, 387-391 (1984).
7. Kavaler, J., Davis, M. M. & Chien, Y.-H. *nature* **310**, 421-423 (1984).
8. Siu, G. *et al. Nature* **311**, 344-349 (1984).
9. Patten, P. *et al. Nature* **312**, 40-46 (1984).
10. Siu, G. *et al. Cell* **37**, 393-401 (1984).
11. Goverman, J *et al. Cell* **40**, 859-867 (1985).
12. Weigert, M. *et al. Nature* **283**, 497-499 (1980).
13. Sakano, H., Kurosawa, Y., Weigert, M. & Tonegawa, S. *Nature* **290**, 562-565 (1981).
14. Kurosawa, Y. *et al. Nature* **290**, 565-570 (1981).
15. Alt, F. & Baltimore, D. *Proc. natn. Acad. Sci. U.S.A.* **79**, 4118-4122 (1982).
16. Kim, S., Davis, M., Sinn, E., Patten, P. & Hood, L. *Cell* **27**, 573-581 (1981).
17. Tonegawa, S. *Nature* **302**, 575-581 (1983).
18. Golub, E. *Cell* **21**, 603-604 (1980).
19. Matzinger, P. & Zamoyska, R. *Nature* **297**, 628 (1982).
20. Doherty, P. C. & Zinkernagel, R. M. *J. exp. Med.* **141**, 502-507 (1975).
21. Thomas, D. W., Yamashita, U. & Shevach, E. M. *J. Immun.* **119**, 223-226 (1977).
22. Hansburg, D., Fairwell, T., Schwarz, R. H. & Appella, E. *J. Immun.* **131**, 319-324 (1983).
23. Hedrick, S. M., Nielsen, E. A., Kavaler, J., Cohen, D. I. & Davis, M. M. *Nature* **308**, 153-158 (1984).
24. Saito, H. *et al. Nature* **309**, 757-762 (1984).
25. Scollay, R. G., Butcher, E. C. & Weissman, I. L. *Eur. J. Immun.* **10**, 210-218 (1980).
26. Malissen, M. *et al. Cell* (in the press).
27. Brodeur, P. H. & Riblet, R. *Eur. J. Immun.* **14**, 922-930 (1984).
28. Cory, S., Tyler, B. M. & Adams, J. M. *J. mol. appl. Genet.* **1**, 103-116 (1981).
29. Brack, C., Hirama, M., Lenhard-Schuller, R. & Tonegawa, S. *Cell* **15**, 1-14 (1978).
30. Hedrick, S. M. *et al. Proc. natn. Acad. Sci. U.S.A.* **82**, 531-535 (1985).
31. Kronenberg, M. *et al. Nature* **313**, 647-653 (1985).
32. Dildrop, R. *Immun. Today* **5**, 85-86 (1984).
33. Crews, S., Griffin, J., Huang, H., Calame, K. & Hood, L. *Cell* **25**, 59-66 (1981).
34. Potter, M. *Adv. Immun.* **25**, 141-211 (1977).
35. Kurosawa, Y. & Tonegawa, S. *J. exp. Med.* **155**, 201-218 (1982).
36. Early, P., Huang, H., Davis, M., Calame, K. & Hood, L. *Cell* **19**, 981-992 (1980).
37. Bernard, O. & Gough, N. M. *Proc. natn. Acad. Sci. U.S.A.* **77**, 3630-3634 (1980).
38. Gough, N. M. & Bernard, O. *Proc. natn. Acad. Sci. U.S.A.* **78**, 509-513 (1981).
39. Yoshikai, Y. *et al. Nature* **312**, 521-524 (1984).
40. Manser, T., Huang, S.-Y. & Gefter, M. L. *Science* **226**, 1283-1288 (1984).
41. Griffith, G. M., Berek, C., Kaartinen, M. & Milstein, C. *Nature* **312**, 271-275 (1984).
42. Augustin, A. A. & Sim, G. K. *Cell* **39**, 5-12 (1984).
43. Wu, T. T. & Kabat, E. A. *J. exp. Med.* **132**, 211-250 (1970).
44. Amzel, L. M. & Poljak, R. J. *A. Rev. Biochem.* **48**, 961-997 (1979).
45. Davies, D. R. & Metzger, H. *A. Rev. Immun.* **1**, 87-117 (1983).
46. Amit, A. G., Mariuzza, R. A., Phillips, S. E. V. & Poljak, R. J. *Nature* **313**, 156-158 (1985).
47. Kabat, E. A., Wu, T. T., Bilofsky, H., Reid-Miller, M. & Perry, H. *Sequences of Immunological Interest* (US Department of Health and Human Services, Washington, DC, 1983).
48. Capra, J. D., Wasserman, R. L. & Kehoe, J. M. *J. exp. Med.* **138**, 410-427 (1973).
49. Chou, P. & Fassman, G. *A. Rev. Biochem.* **47**, 251-276 (1978).
50. Kyte, J. & Doolittle, R. F. *J. molec. Biol.* **157**, 105-132 (1982).
51. Poljak, R. J., Amzel, L. M., Chen, B. L., Phizackerley, R. P. & Saul, F. *Proc. natn. Acad. Sci. U.S.A.* **71**, 3440-3444 (1974).
52. Saul, F. A., Amzel, L. M. & Doljak, R. J. *J. biol. Chem.* **253**, 585-597 (1978).
53. Wylie, D. E., Sherman, L. A. & Klinman, N. R. *J. exp. Med.* **155**, 403-414 (1982).
54. Rodwell, J., Gearhart, P. & Karush, F. *J. Immun.* **130**, 313-316 (1983).
55. Mitchell, G. F. & Miller, J. F. A. P. *J. exp. Med.* **128**, 821-837 (1968).
56. Huynh, T. V., Young, R. A. & Davis, R. W. in *DNA Cloning: A Practical Approach* (ed. Glover, D.) (IRL, Oxford, 1984).
57. Sanger, F., Nicklen, S. & Coulsen, A. *Proc. natn. Acad. Sci. U.S.A.* **74**, 5463-5467 (1977).
58. Strauss, E. C., Kobori, J. A., Siu, G. & Hood, L. E. *Analyt. Biochem.* (in the press).

Cell, Vol. 60, 929–939, March 23, 1990.

Chimeric Immunoglobulin–T Cell Receptor Proteins Form Functional Receptors: Implications for T Cell Receptor Complex Formation and Activation

Joan Goverman,* Stephen M. Gomez,†
Kathleen D. Segesman,* Tim Hunkapiller,*
Walter E. Laug,† and Leroy Hood*
*Division of Biology
California Institute of Technology
Pasadena, California 91125
†Department of Hematology-Oncology
Childrens Hospital of Los Angeles
Los Angeles, California 90027

Summary

We constructed chimeric receptor chains in which an immunoglobulin heavy chain variable region (V_H) from a phosphorylcholine-specific antibody is substituted for T cell receptor (Tcr) α and β V regions. We demonstrate that the V_H region joined to either the C_α or the C_β region can form stable chimeric proteins in EL4 T cells. Both chimeric receptor chains associate with CD3 polypeptides in functional receptor complexes and respond to phosphorylcholine coupled to Sepharose beads. The V_H–C_α chimeric chain associates with the EL4 β chain, while the V_H–C_β chimeric protein appears to form either a homodimer or a heterodimer with the native EL4 β chain. Thus, functional receptor complexes can be formed using two C_β regions, and the C_α region may not be required for CD3 association and surface expression of Tcr complexes.

Introduction

Two types of lymphocytes are responsible for the recognition of foreign antigen in the immune response. The cell surface receptors of T cells recognize antigenic peptides only when they are complexed with polypeptides encoded in the major histocompatibility complex (MHC) and expressed on the surface of antigen-presenting cells. Thus, engaging a Tcr requires the formation of a trimolecular complex in which all components are displayed on a cell surface (Schwartz, 1984). B cells, in contrast, use cell surface immunoglobulin to recognize the three-dimensional conformation of antigens independent of a cellular or MHC-associated context (Davies et al., 1988). Other interactions between T cells and target cells also contribute toward limiting T cell responses to cellularly presented antigen. The T cell accessory molecules CD4 and CD8 appear to stabilize cell–cell interactions by interacting with nonpolymorphic determinants on MHC molecules and may also participate in transmitting regulatory signals (Parnes, 1989). Also, activation of resting T cells appears to require two signals (Chiller et al., 1982; Mizel, 1982). One signal consists of engaging the Tcr, and the other is transmitted by lymphokines produced by the antigen-presenting cell. Thus, to receive a signal required for proliferation, T cells must be proximal to another cell that is producing that signal.

Both B and T cell receptors are disulfide-linked heterodimers consisting of light (L) and heavy (H) chains for immunoglobulin and α and β chains for the predominant Tcrs. Each chain is divided into variable (V) and constant (C) regions. The V regions of both chains associate to form the antigen-binding domain, and the C regions link the receptor to the cell surface and, in the case of Tcr, to other cell surface accessory molecules (Davies and Metzger, 1983; Davis, 1985; Kronenberg et al., 1986). The immunoglobulin C regions are associated with the effector function of the molecule. For cell surface expression, the Tcr must be associated with a complex of polypeptides called CD3 (Clevers et al., 1988). These interactions appear to involve the C regions and mediate signal transduction between the receptor and cell (Weiss et al., 1986). A second type of Tcr associated with CD3 is composed of γ and δ chains and is expressed on immature thymocytes and 1%–2% of peripheral T cells (Saito et al., 1984; Chien et al., 1987; Brenner et al., 1986). While δ and γ chains exhibit sequence similarity to α and β chains, respectively, they express a more limited V region repertoire and do not appear to play a major role in the general MHC-restricted response to antigens (Strominger, 1989).

The V regions of the H and β chains are formed from separate V, diversity (D), and joining (J) gene segments that rearrange together to form a functional V gene. The L and α V regions are formed by rearranging only V and J gene segments. Predicted secondary and tertiary structures based on sequence comparisons of these V regions indicate a fundamental similarity between the antigen-binding domains of the two receptors (Goverman et al., 1986; Novotný et al., 1986). Thus, no structural features are apparent in the V_α and V_β regions of the Tcr that could account for its unique properties of antigen recognition in relation to those of immunoglobulin.

We are interested both in defining the molecular interactions between T cells and target cells, which signal activation, and in identifying features of Tcr chains that mediate their interaction with one another and CD3 components. Our approach is to construct chimeric immunoglobulin–Tcr genes in which the V region genes of the Tcr are replaced with those of an immunoglobulin of known specificity. Transfecting these chimeric receptor genes into T cells should confer a specificity for an antigen that can be easily manipulated and is independent of MHC determinants. The first step in setting up this system is to determine whether there is a unique character to Tcr V regions not shared by immunoglobulin V regions. To examine this question directly, we determined the ability of individual chimeric receptor chains to pair with native Tcr chains and form receptor complexes.

We report here the characterization of chimeric receptor genes encoding an immunoglobulin V_H region that is joined to either the C_α or C_β regions of a Tcr. The V_H re-

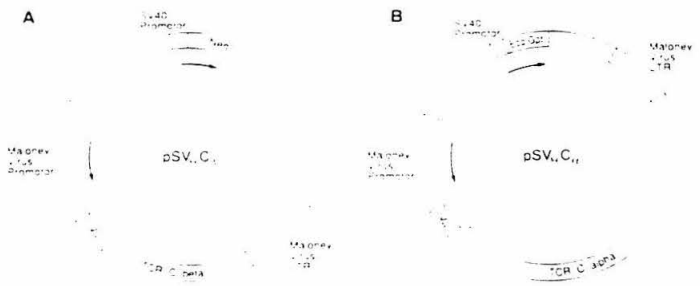


Figure 1. Schematic of the Plasmids Containing the Chimeric V_H–C_β and V_H–C_α Genes

The direction of transcription is indicated for each construct. Both chimeric genes are flanked by Maloney LTR sequences; the 5′ LTR is used as the promoter.

(A) The V_H–C_β construct subcloned into pSV2-neo. The V_H-containing fragment begins 65 bp 5′ and ends 70 bp 3′ of V coding sequences. These sequences do not include an intact immunoglobulin promoter. The C_β2 gene is on a HindIII fragment containing 700 bp of 5′ sequence and 500 bp of 3′ sequence.

(B) The V_H–C_α construct subcloned into pSV2-gpt. The V_H sequences are as in (A); the C_α gene is on a BamHI fragment containing 600 bp of 5′ sequence and 5 kb of 3′ sequence.

gion of the antibody S107 comprises much of the binding site for its hapten, phosphorylcholine (PC), and may be largely responsible for determining antigen specificity (Segal et al., 1974). PC specificity is retained when this V_H region is paired with three very different V_L regions (Rodwell et al., 1983). Thus, expression of a receptor chain containing this V_H gene can confer antigen specificity when paired with a variety of other receptor chains. Both chimeric receptor proteins are expressed on the cell surface, are associated with CD3, and form functional receptors that respond to antigen stimulation.

Results

Design of Chimeric Receptor Gene Constructs

The immunoglobulin V gene used in our chimeric receptor constructs was derived from the rearranged V_H gene of the PC-specific mouse B cell myeloma S107 (Rudikoff et al., 1979; Crews et al., 1981). The PC ligand can be used in soluble form as well as coupled to solid supports, various proteins, or cells. A genomic clone of the V_H gene and anti-idiotype antibodies, specific for several distinct epitopes on the V_H region independent of association with a particular V_L region, were available (Desaymard et al., 1984).

In our constructs, we used a 700 base pair (bp) fragment that contained the rearranged V_H gene without the promoter sequences (Clarke et al., 1982). This fragment was subcloned into an expression vector containing the murine Maloney retroviral promoter sequences. For the V_H–C_β construct, a 2.9 kb genomic fragment containing the mouse C_β2 gene was subcloned into this plasmid 3′ to the V_H gene (Figure 1A). Thus, the chimeric gene retains the appropriate splicing signals between the V and C genes, and its expression is controlled by a viral promoter that is active in T cells but independent of Tcr-specific enhancer sequences. For the V_H–C_α construct, a 9 kb genomic fragment containing the mouse C_α gene was subcloned 3′ to the V_H gene in the same expression vector (Figure 1B). The V_H–C_β expression construct was then subcloned into pSV2-neo and the V_H–C_α expression construct was subcloned into pSV2-gpt for transfection into T cells (Mulligan and Berg, 1980; Southern and Berg, 1982).

Chimeric Receptor Proteins Are Transcribed and Expressed on the Cell Surface

Each of these chimeric receptor genes was transfected into the mouse T cell lymphoma EL4, and drug-resistant clones were isolated (Herberman, 1972; Shevach et al., 1972). Hybridization of Northern blots with a V_H gene–specific probe indicated that transcription of the V_H–C_β and V_H–C_α genes generates the appropriate size messages (Figure 2). EL4 expresses its own Tcr, and the corresponding transcripts are shorter than the chimeric receptor transcripts. Therefore, we were able to hybridize Northern blots with C_β and C_α probes to compare the

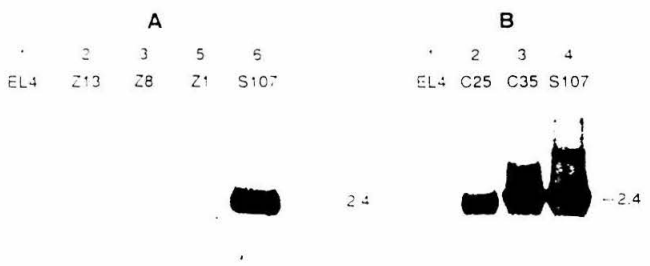


Figure 2. Northern Blot Analyses of RNA from EL4 Clones Transfected with the Chimeric Receptor Genes

Poly(A)$^+$ RNAs were detected on Northern filters by hybridization with a 700 bp V_H probe labeled by primer extension as described in Goverman et al. (1985).

(A) V_H–C_β-transfected clones: EL4 (lane 1). V_H–C_β-transfected clones (lanes 2–5). S107 (lane 6). The autoradiogram is a 4 hr exposure.

(B) V_H–C_α-transfected clones: EL4 (lane 1). V_H–C_α-transfected clones (lanes 2 and 3). S107 (lane 4). The autoradiogram is a 20 hr exposure.

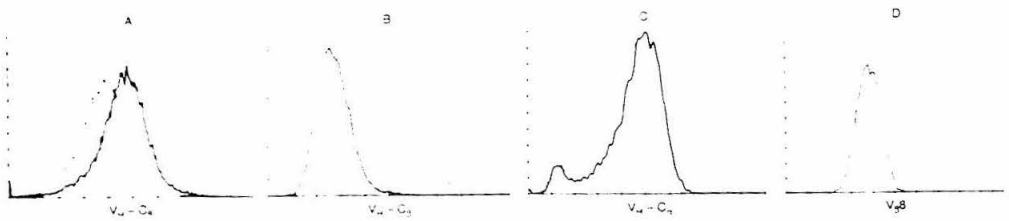


Figure 3. Cell Surface Expression of the Chimeric V_H–C_γ and V_H–C_α Gene Products

(A) V_H–C_γ transfected clone Z1 stained with FITC-conjugated anti-rat IgG (broken line) or anti-V_H antibody 68.3 followed by FITC-conjugated anti-rat IgG (solid line).

(B) Clone Z1 stained as in (A) except that anti-V_H antibody 85.4 was used instead of 68.3 (solid line).

(C) V_H–C_α-transfected clone C25 stained with FITC-conjugated anti-rat IgG (broken line) or anti-V_H antibody 85.4 followed by FITC-conjugated anti-rat IgG (solid line).

(D) A T cell hybridoma expressing an endogenous Tcr containing $V_\beta 8.1$ stained with FITC-conjugated anti-mouse IgG (broken line) or the V 8-specific antibody F23.1 (solid line) (Staerz et al., 1985).

level of chimeric gene expression with that of endogenous Tcr gene expression. The level of mRNA expressed from both constructs was at least 5-fold higher in all clones tested (data not shown).

Several mRNA-producing clones were characterized by flow cytometry to determine whether the chimeric receptor proteins are expressed on the cell surface. Two monoclonal anti-idiotype antibodies recognizing different S107 V_H epitopes, 85.4 and 68.3, stained the V_H–C_γ-transfected clone Z1 (Figures 3A and 3B; Desaymard, 1984). These results indicate that the V_H–C_γ protein is expressed on the cell surface and must generally maintain the native conformation of the V_H region. Flow cytometry analyses also revealed cell surface expression of the chimeric V_H–C_α protein on the V_H–C_α-transfected clone C25 (Figure 3C). Interestingly, only the 85.4 antibody stained the cell surface, suggesting that either the conformation of the V_H region was altered in this chimeric protein or that the epitope recognized by 68.3 may be inaccessible due to association with a different V region.

The Chimeric Receptor Proteins Associate with CD3 Polypeptides

To determine whether the chimeric receptor proteins are associated in complexes with CD3 polypeptides, we carried out two-dimensional gel analyses on surface-labeled proteins immunoprecipitated with an anti-CD3 antibody, 2C11 (Leo et al., 1987). Figure 4A shows the resolution of the endogenous α and β chains obtained by immunoprecipitating surface-labeled proteins of EL4 with 2C11. Similar experiments with the V_H–C_γ transfectant revealed a large increase in proteins migrating with a pI and molecular weight similar to the endogenous β chain (Figure 4B). Similarly, an increase in protein expression in the position corresponding to the α chain was observed in the V_H–C_α-expressing transfectant (Figure 4C). Thus, immunoprecipitation with 2C11 indicated that the chimeric receptor proteins are assembled into a Tcr-like complex with CD3 polypeptides. These results demonstrated that replacement of either Tcr V region with an immunoglobulin V_H region does not affect the ability of CD3 polypeptides to associate with the chimeric Tcr chains.

Characterization of the Disulfide-Linked Pairing of the Chimeric Receptor Proteins

To determine whether the chimeric receptor proteins paired with the endogenous receptor chains, we carried

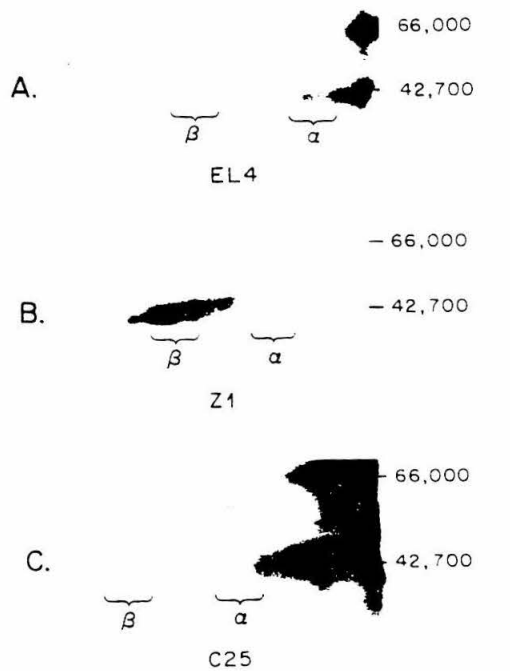


Figure 4. Two-Dimensional Gel Electrophoresis of ^{125}I-Labeled Cell Surface Proteins Immunoprecipitated with the Anti-CD3 Antibody 2C11

The first dimension was isoelectric focusing under reducing conditions. This was followed by SDS–PAGE on a 10% polyacrylamide gel. (A) EL4. (B) V_H–C_γ-expressing clone Z1. (C) V_H–C_α-expressing clone C25. The brackets indicate the positions of the α and β polypeptides migrating at about pH 4.5 and 7.0, respectively. The horizontal pH gradient was determined in the first dimension using Biorad isoelectric focusing standards.

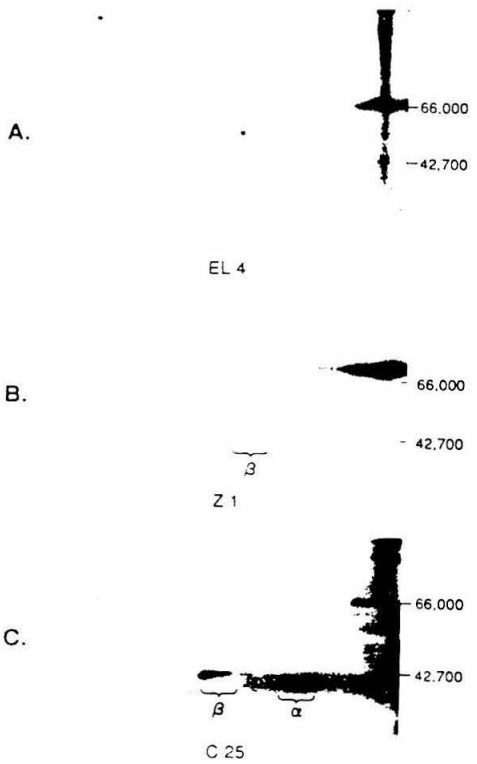


Figure 5. Two-Dimensional Gel Electrophoresis of ^{125}I-Labeled Cell Surface Proteins Immunoprecipitated with the Anti-V_H Antibody 85.4

Conditions were as described for Figure 4. The 85.4 antibody was bound to goat anti-rat-coated Staph A prior to incubation with whole cell lysates. (A) EL4. (B) clone Z1. (C) clone C25.

out two-dimensional gel analysis on surface-labeled proteins immunoprecipitated with the anti-V_H antibody 85.4. If the chimeric receptor proteins are paired with heterologous EL4 chains, immunoprecipitation with 85.4 should precipitate proteins that migrate at positions corresponding to both the α and β chains. This was the case for the transfectant expressing the V_H–C_α protein, C25 (Figure 5C). Thus, this chimeric protein appears to pair with the EL4 β chain. Similar analyses with the V_H–C_β-expressing transfectant, however, did not detect any α chains in the receptor complexes formed with this chimeric protein (Figure 5B).

We also analyzed the two transfectants by diagonal gel electrophoresis of surface-labeled proteins immunoprecipitated with the anti-V_H antibody 85.4 or anti-CD3 antibody 2C11. Because only the second dimension of these gels is run under reducing conditions, only disulfide-linked polypeptides migrate off the diagonal line. The dissociated α and β Tcr chains migrated off the diagonal at different molecular weights, generating two distinguishable spots. Figure 6A shows the resolution of the EL4 α and β polypeptides immunoprecipitated with 2C11. Immunoprecipitation with 2C11 of labeled proteins from the transfectants will precipitate both the normal, EL4 Tcr complexes as well as the chimeric protein–containing receptor complexes. The replacement of the Tcr V regions with the immunoglobulin V_H region is not expected to alter significantly their molecular weights from the native receptor chains. For both transfectants, two polypeptides migrated off the diagonal at the positions of α and β chains (Figures 6B and 6C). The same analyses were carried out using the anti-V_H antibody 85.4. Two polypeptides migrated off the diagonal for the V_H–C_α-expressing transfectants, but only one spot appeared off the diagonal for the V_H–C_β-expressing transfectant (Figures 6E and 6F). These experiments were repeated multiple times with the same results. These data suggest that the V_H–C_α chimeric proteins pair with the EL4 β chain, while the V_H–C_β chimeric proteins pair primarily with another C_β-containing receptor chain. Because these gels do not resolve β chains containing different V regions, our data do not distinguish whether this single spot represents a homodimer (i.e., two V_H–C_β chains), a heterodimer of chimeric and normal β chains, or both.

To confirm our result of generating C_β dimers, we transfected the V_H–C_β gene into a mutant of the human T cell line Jurkat (Saito et al., 1987). This mutant does not express a Tcr or CD3 polypeptides on the cell surface, because a functional α chain transcript is not produced. Previous experiments showed that transfection of this mutant with a murine α chain gene restored surface CD3 expression (Saito et al., 1987). We characterized 18 clones expressing transcripts hybridizing to the V_H probe by flow cytometry. Nine clones stained with a human anti-CD3 antibody, OKT3 (Figure 7). This result indicates that C_β dimers formed in this system as well and are associated with CD3 polypeptides.

V_H-Containing Chimeric Receptor Proteins Assemble into Functional Tcr Complexes

We tested the V_H–C_β- and V_H–C_α-expressing transfectants for the ability to produce IL-2 when stimulated with a derivative of PC, p-aminophenyl PC (PC*). We used PC* because the affinity of S107 for PC coupled to a phenyl group is increased about 5-fold in relation to PC alone, and this derivative is the form of PC coupled to Sepharose beads (Rodwell et al., 1983). Neither transfectant produced IL-2 when incubated with soluble PC* (Figure 8). However, both transfectants responded to PC* coupled to Sepharose beads. A low level of lymphokine release was also observed when the transfectants were stimulated with Sepharose beads alone. This response may result from interaction with the cross-linked carbohydrate groups that compose Sepharose beads as the S107 antibody binds specifically to bacterial cell membranes containing PC linked to carbohydrate groups (Kohler, 1975). The lack of IL-2 production by EL4 in response to all stimuli shows that the response to PC*-coupled Sepharose and Sepharose beads was completely transfectant specific. These results indicate that the V_H–C_α chain forms a functional complex with the native EL4 β chain. The V_H–C_β chimeric protein paired with another C_β-containing chain also

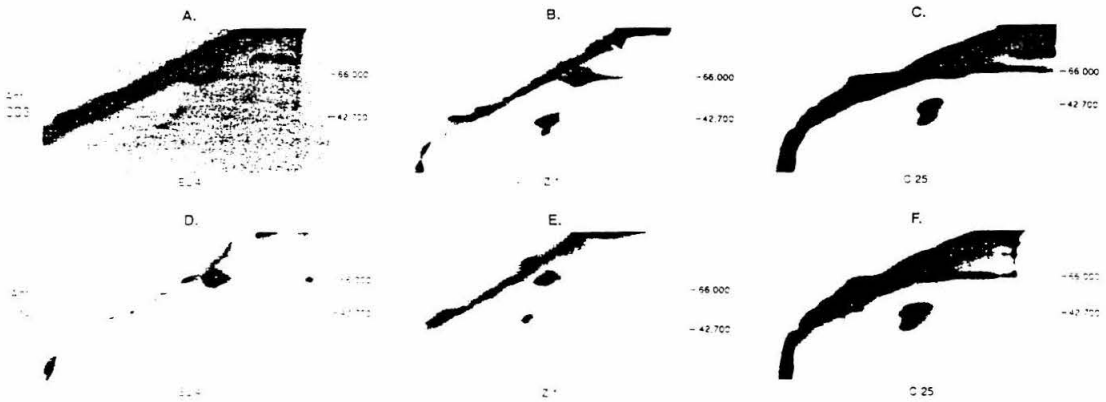


Figure 6. Two-Dimensional (Diagonal) SDS–PAGE Analysis of Chimeric Receptor-Expressing Transfectants

^{125}I-labeled cell surface proteins were immunoprecipitated with either 2C11 or 85.4. SDS–PAGE in the first dimension was on 7.5% polyacrylamide gels under nonreducing conditions. The second dimension was on 10% polyacrylamide gels under reducing conditions. (A) EL4 immunoprecipitated with 2C11. (B) Z1 immunoprecipitated with 2C11. (C) C25 immunoprecipitated with 2C11. (D) EL4 immunoprecipitated with 85.4. (E) Z1 immunoprecipitated with 85.4. (F) C25 immunoprecipitated with 85.4.

forms a functional complex. It is possible, however, that any functional activity in this transfectant is the result of a small number of V_H–C_μ/V_λ–C_λ heterodimers not detected by our gel system.

Figure 8 also shows that the response to PC*-coupled Sepharose beads is not inhibited by pre- or coincubation with soluble PC*. This suggests that the affinity of the V_H-containing receptors for PC*-coupled Sepharose is higher than for soluble PC*. We do not know the affinity of the V_H-containing Tcrs for PC*. However, because the pairing of this V_H gene segment with three unrelated L chains generates antibodies with affinities for PC* that range from 5–25 × 10^{-5}, we expect the affinity of our chimeric receptor to fall within or below this general range (Rodwell et al., 1983). This implies that if the affinity of a Tcr for its

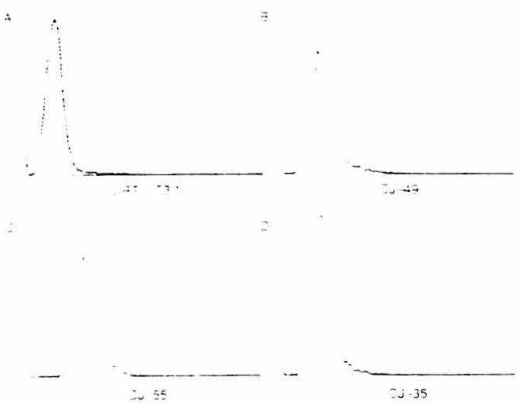


Figure 7. Flow Cytometry Analysis of V_H–C Transfected Clones Derived from the α Negative Mutant of Jurkat J RT-T3.1

The parent cell line and transfected clones were stained with the anti-human CD3 antibody OKT3 to detect surface expression of the CD3 complex. The broken lines represent staining with FITC-conjugated anti-mouse antibody only and the solid lines represent staining with OKT3 followed by FITC-conjugated anti-mouse. (A) J RT-T3.1. (B) CJ-49. (C) CJ-55. (D) CJ-35.

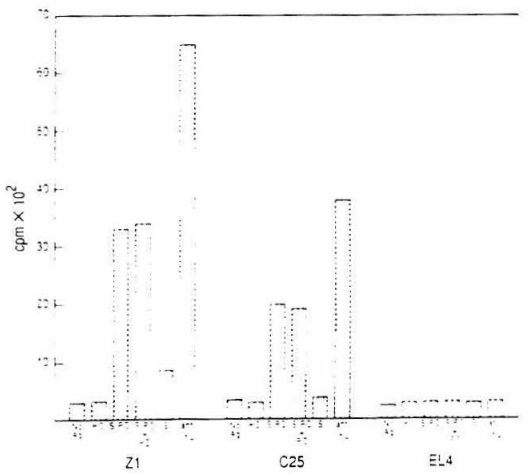


Figure 8. Lymphokine Response of Transfectants Expressing the Chimeric Receptor Gene Products

Clones Z1, C25, and EL4 were incubated for 24 hr with various stimuli, and the supernatants were then removed and tested for IL-2 production by the ability to support growth of HT-2 cells. Proliferation of the HT-2 cells was assayed by measuring the incorporation of [^{3}H]thymidine 24 hr after transfer of the supernatants. The reagents used to stimulate the cells were: No Ag, culture medium; PC, p-aminophenyl phosphorylcholine; S-PC, Sepharose beads coupled to p-aminophenyl phosphorylcholine; S, Sepharose beads (6B-CL); anti-V_H, antibody 85.4.

ligand complex is on the order of 10^{-5} or less, it would not be activated by soluble ligand. Thus, either T cells require higher affinity interactions to be activated by soluble antigen or presentation of antigen in a solid-phase context changes the affinity requirements. A similar conclusion was reached by Becker et al. (1989) studying bulk T cell lines expressing chimeric V_H–C_α receptors established from transgenic mice. However, the affinities of these chimeric protein–containing receptors could not be estimated more accurately than within a range of 10^{-5}–10^{-9}.

Both transfectants were also activated by incubation with the soluble anti-V_H antibody 85.4 without secondary cross-linking with anti-IgG. These results confirmed the ability of V_H–C_α-containing receptors to transduce a signal for IL-2 production and demonstrated that a receptor complex with two C_β regions can also trigger activation.

Discussion

We characterized Tcr incorporating a chimeric protein in which the V_H region from an antibody specific for PC is paired with either Tcr C region. These receptors represent functional complexes, because they consist of disulfide-linked dimers, associate with CD3 polypeptides, and respond specifically to antigen stimulation. As observed for anti-PC antibodies, the V_H region appears to confer a specificity for PC. Thus, we generated T cells specific for an antigen that can be presented in either a soluble or solid-phase context. Analysis of how these chimeric proteins assemble into functional complexes allowed us to probe the interactions that occur between the Tcr α and β chains as well as between these chains and the CD3 polypeptides. Ultimately, this system should allow us to investigate the requirements for antigen presentation in triggering activation and the role of accessory molecules independent of Tcr–MHC interactions.

Functional Implications

Kuwana et al. (1987) described receptors in which an entire Tcr V domain was replaced by an immunoglobulin V domain. V genes from a PC-specific antibody were used in these constructs, including the same germline V_H gene segment as that of S107. Ca^{2+} flux, an early event in T cell activation, was observed when EL4 cells transfected with the constructs were exposed to PC-expressing bacteria. However, neither protein expression of the chimeric chains nor the response of single chain transfectants was analyzed. Our results indicate that the V_H-containing chimeric chains can associate with other receptor chains not expressing a V_L region and still confer specificity for PC. Thus, confirmation that a T cell V domain can be replaced with an immunoglobulin V domain requires further characterization.

Our work and that of Becker et al. (1989) showed that a V_H region can functionally replace a V_α region by associating with a V_β region to form receptors capable of triggering activation. We also showed that the V_H region can form functional receptors when joined to the C_β region. Comparison of these different transfectants resulted in interesting structural and functional observations. The V_H–C_β chimeric protein appears to pair with another C_β-containing polypeptide. The occurrence of C_β dimers was confirmed by the appearance of surface-expressed CD3 in an α-negative T cell mutant transfected with the V_H–C_β construct. The ability of the transfectant Z1 to respond to both PC*-coupled Sepharose and an anti-V_H antibody implies that a functional Tcr can be formed with two C_β regions.

Our transfectants responded to PC*-coupled Sepharose but did not respond to soluble PC*. Three possible explanations could account for this observation. First, there may be an increased affinity of the chimeric receptors for PC*-coupled Sepharose such that the receptor–ligand interaction is now sufficient for signal transduction. Our data suggest a higher affinity of the V_H-containing receptors for PC*-coupled Sepharose, because soluble PC* does not inhibit the response to PC*-coupled Sepharose. It is possible, in fact, that soluble PC* does not bind to the chimeric receptors at all. Because the T cells are stimulated over a 24 hr period, however, even a small difference in affinity for PC* vs. PC*-coupled Sepharose could result in the lack of inhibition. Second, the avidity of the T cells for the Sepharose beads may be increased by making the ligand multimeric. Increased avidity could lead to a receptor occupancy that is not achieved in the soluble form and that is sufficient for activation. Third, a multimeric, solid-phase ligand may be required to activate T cells. Consistent with this hypothesis, studies using MHC molecules and alloreactive T cells demonstrated that the MHC antigen had to be integral to planar membranes or coupled to solid supports to initiate activation (Kappler et al., 1983; Meuer et al., 1983; Herrmann and Mescher, 1986; McCluskey et al., 1988).

Our results do not clearly distinguish between the explanations described above. However, recent experiments by Schneck et al. (1989) support the requirement for a solid-phase ligand. They demonstrated that a soluble MHC molecule was unable to stimulate a T cell hybridoma but could inhibit T cell stimulation by cells presenting a related MHC molecule. This implies that receptor occupancy must occur with soluble MHC molecules but that this does not result in activation. The conclusion drawn from these experiments, as well as from those using immobilized ligands, is that Tcr cross-linking is required for signal transduction.

Other experiments using antigen–MHC complexes in planar membranes indicated that maximum IL-2 production by T cell hybridomas was obtained when the complexes were about 200 Å apart, and no IL-2 production was observed at a distance of 800 Å or more (Watts, 1988). However, the low concentration of relevant antigen–MHC ligands under physiological conditions implies that the complexes are sufficiently far enough apart such that cross-linking of Tcr would occur only by passive diffusion (Allen et al., 1987; Kourilsky et al., 1987). This passive diffusion should occur with Tcr occupied by soluble ligand as well, and yet this does not result in activation. We suggest instead that the requirement for solid-phase ligand may involve a change in the state of the receptor (or the complex of molecules associated with it) that occurs as a

consequence of engaging a ligand that is fixed to another surface. Cross-linking is not necessarily required in this scenario; it is only required that some change in the state of the receptor complex occurs.

A paradigm for changes in membrane proteins that result from mechanical interaction can be found in stretch-activated ion channels. These channels are activated by physical interactions that can result from cell–cell contact (Kullberg, 1987; Lansman et al., 1987; Arcangeli et al., 1989; Morris and Sigurdson, 1989). Similarly, interaction with a surface-bound ligand may induce a change in the Tcr. This change is not likely to confer the ability to transduce a signal directly, since activation is not observed when T cells are stimulated with MHC molcules anchored in a planar membrane at a distance far enough apart to prevent cross-linking. We suggest instead that mechanical tension between receptor complexes and solid-phase ligand induces a change in the receptor complex that alters the affinity of ligand-bound receptors for one another or for other molecules. Polypeptide associations may then occur that initiate activation. The coclustering of CD4 and Tcr that occurs upon ligand binding as well as the association of other polypeptides with the Tcr upon activation are consistent with this idea (Fraser et al., 1989; Rojo et al., 1989). Our data and that of others using soluble bivalent antibodies argue that extensive cross-linking is not required (Haskins et al., 1984; Hua et al., 1986; Lanier et al., 1986; Rojo and Janeway, 1988). These data also argue that the ability of solid-phase ligand to activate T cells is not merely the result of concentrating into one region of the membrane Tcr that has activated the signal transduction pathway. Thus, given the monovalent nature of the Tcr and its ligand, activation by simple receptor cross-linking may not represent the normal physiological mechanism of activation.

Structural Implications

Our results demonstrate the ability of an immunoglobulin V_H region to substitute for the Tcr V_α region to generate a functional receptor. In a reciprocal system, chimeric chains with a V_α region joined to an immunoglobulin C_H region were able to associate with an immunoglobulin L chain and form a secretable, disulfide-linked heterodimer (Gascoigne et al., 1987). Thus, the V_α and V_H regions appear to be able to substitute for one another at least at the level of stable dimer formation, indicating that they share similar structural properties.

In contrast to its ability to substitute for V_α, our experiments indicate that the V_H region did not substitute well for V_β in pairing with the EL4 α chain. However, analyses of V region sequences indicate that the V_H region is structurally no more similar to V_α than it is to V_β (Hunkapiller and Hood, 1989). Thus, we would predict that, if the interaction of the V_H–C_α chain with the native α chain is inhibited because the two V regions cannot associate, this may be peculiar to this particular pair of V regions rather than a general inability of V_H and V_α regions to associate. In this regard, different pairs of V_H and V_L regions can have significantly different rates of association in vitro, and there is no preferential association of autologous vs. heterologous chains (Hamel et al., 1987).

Another possible explanation for the lack of association between the chimeric V_H–C_β protein and the native α chain is the incompatibility of the overall structural geometries of the two chains. For example, the angle formed between the immunoglobulin V_L and C_L regions, governed mostly by the junctional J and C region residues, is different from the angle formed between the V_H and C_H regions (Amzel and Poljak, 1979). X-ray crystalographic studies on naturally occurring L chain dimers showed that this geometry can be flexible in the L chain, because one chain in the dimer assumes the angle typical of the H chain. Analogous H chain dimers have not been found, perhaps indicating the inflexibility of the H chain joint. Tcr chains may also have distinct V-to-C angles that may or may not be flexible. One interpretation of our results is that the V_H–C_α chain has a geometry like that of a native α chain but that the V_H–C_β chain geometry cannot assume the geometry of the native β chain. This would imply that the α chain is not flexible enough to accommodate the V_H–C_β geometry. To test these ideas, we are characterizing other V_H–C_β chimeric proteins in different recipient T cells to determine whether the inability of these proteins to pair with native α chains is a generalizable result.

Although our results demonstrate that dimers containing two C_β regions can form cell surface receptors, Tcr β chain dimers have not been observed in T cell mutants deficient in α chain expression (Saito et al., 1987). The lack of β chain dimers could reflect either an inability of two V_β regions to associate or a requirement for different, inflexible geometries in a two chain receptor. Recent studies characterizing Tcr polypeptides in which the transmembrane sequences were substituted with sequences that encode a phosphoinositol-membrane linkage demonstrated that a mixture of homo- and heterodimers of both α and β chains appeared on the surface of COS cells (Davis et al., 1989). However, these dimers are expressed on the surface without CD3 polypeptides, and it is unclear that their interaction resembles the chain interactions in Tcr–CD3 complexes. We hypothesized that the V_H region assumes an angle between both the C_α and C_β regions that is similar to the angle between the V_α and C_α regions. This idea predicts that the V_H–C_α would pair with V_β–C_β and that the V_H–C_β would also pair with V_β–C_β and not with V_α–C_α. This is consistent with our data and implies that the C_β dimer we observe is a heterodimer consisting of V_H–C_β and V_β–C_β.

Recently, functional rearrangements that form chimeric genes encoding V_δ–C_γ and V_γ–C_δ have been detected in lymphoid tissues, presumably arising from chromosomal translocations (Tycko et al., 1989). Because of the sequence similarities between α and δ chains and between γ and β chains, it would be interesting to see whether a V_δ–C_γ protein would combine with a native δ chain or, like our V_H–C_β, generate C_γ-containing dimers. These chimeric proteins differ from the ones described here, however, because the J regions belong to the same family as the C region. It is possible that the V gene segments in different receptor families encode structurally interchange-

able protein sequences and that it is the overall chain geometries, as governed mostly by J and C gene segment sequences, that determine possible dimer formation. Thus, these chains may be more likely to retain the geometry characteristic of their native C region–containing chain.

Our results suggest that interaction between C_α and CD3 is not required for surface expression of a Tcr receptor, but that a dimer of some form is required. Chemical cross-linking experiments have indicated that the C_β region associates with the CD3 γ polypeptide (Brenner et al., 1985). Although the precise residues involved in these interactions are not known, charged residues in the transmembrane sequences may play a role. The occurrence of C_β dimers suggests that either one of the C_β regions in the dimer must substitute for the C_α region in CD3 interactions or that the C_α region does not play a required role in normal receptor complexes. Biochemical analyses of some human T cell clones suggesting that γ–γ homodimers may be expressed on the surface without coexpression of the α equivalent δ chain support the idea that C_α may not be needed for receptor–complex formation (Alarcon et al., 1987; Borst et al., 1987).

The C_α region has several interesting structural features. It contains a sequence homologous to all other C regions, referred to as the immunoglobulin homology unit (Hill et al., 1966). In relation to most other immunoglobulin C region homology units however, the C_α homology unit lacks a sequence of about ten residues that contains an otherwise invariant tryptophan residue. This residue is thought to be important in stabilizing the conserved disulfide bond of the immunoglobulin C homology unit (Amzel and Poljak, 1979).

The portion of C_α usually identified as the transmembrane region is also unusual. While transmembrane sequences are typically thought of as having a stretch of about 21 hydrophobic residues, individual charged residues are occasionally included. However, a sequence is not usually predicted to be transmembrane without at least 11 contiguous hydrophobic residues and usually more than 18 (Eisenberg et al., 1984). This region of C_α contains two charged residues positioned such that there are no stretches of hydrophobic amino acid sequence longer than five residues. In some transmembrane sequences, single charged residues can be "hidden" by interaction with another transmembrane polypeptide. Such interaction is usually predicted to result from the amphipathic properties of both transmembrane segments. If this region in C_α assumes an α-helical structure similar to that predicted for other transmembrane sequences, simple modeling indicates that the positively charged residues are distributed on opposite sides such that an amphipathic helix would not be formed (data not shown). Thus, C_α does not fit the standard model of a transmembrane protein and may not, therefore, be transmembrane.

Alternatively, C_α may interact with at least two components of CD3 to shield each charged residue. The difficulty in rationalizing how C_β or C_γ dimers, whose transmembrane segments contain only one charged residue, could interact with CD3 components in the same way makes this possibility seem less likely. The Tcr δ chain, which can also combine with the β chain if no α chain is present, shares all of these unusual features (Hochstenbach and Brenner, 1989). In contrast, the β chain meets the several criteria for a transmembrane portion even with the presence of one charged residue. Analyses predict that it does form an amphipathic helix and traverses the membrane in association with some other transmembrane segment (after Eisenberg et al., 1984; data not shown). If the other transmembrane segment is a CD3 polypeptide rather than the α chain, then Tcr dimers may be analogous to H and L chain dimers in that they have only one transmembrane chain.

In summary, we demonstrate that chimeric V_H–C_α and V_H–C_β polypeptides form α–β and β–β dimers, respectively, that associate with CD3 polypeptides and are capable of activating T cells in response to polymeric antigen. Moreover, our observations suggest that the α chain may not be required for assembly of the CD3 chains into the T cell receptor molecular complex. Therefore, the functional role of the C_α region in Tcr complexes and signal transduction remains an interesting area for future investigation.

Experimental Procedures

Transfection

EL4 cells were transfected by electroporation using a Biorad Gene Pulser set at 260 V and 960 μF. Cells (1×10^7) were incubated on ice for 10 min in 1 ml of phosphate-buffered saline (PBS)/1% DMSO prior to transfection. After pulsing, cells were incubated for 10 min on ice and plated at 4×10^6 cells/ml in 100 mm dishes in DMEM supplemented with 10% fetal calf serum (FCS), 1 mM pyruvate, 2 mM glutamine, 50 U/ml penicillin, 50 μg/ml streptomycin, and 50 μM 2-mercaptoethanol. The medium was changed after 24 hr, and the selective medium was added after 48 hr. The medium was changed every 3 days afterward. For the pSV2-neo construct, Geneticin (Gibco) was added at 800 μg/ml. For pSV2-gpt, HAT medium (Oi and Herzenberg, 1980), 250 μg/ml xanthine, and 6 μg/ml mycophenolic acid (Gibco) was added. Colonies were visible at 3 weeks and subcloned at limiting dilution. Jurkat cells were transfected as described above, except that the cells were transferred after pulsing to a T75 tissue culture flask containing 25 ml of growth medium (RPMI supplemented with 5% FCS, 1 mM pyruvate, 2 mM glutamine, 50 U/ml penicillin, and 50 μg/ml streptomycin). The cells were incubated for 3 days and then plated in 24-well plates at 7.5×10^4 cells/ml in medium supplemented with 2 mg/ml Geneticin. Colonies were visible at 2 weeks and subcloned at limiting dilution.

Flow Cytometry Analysis

Lymphocyte suspensions containing 2×10^6 cells/100 μl in PBS/0.1% FCS were incubated for 45 min on ice with 100 μl of antibody at various dilutions. Antibody dilutions were: 85.4 and 68.3, 1:1 dilution in FCS of culture supernatants from the rat hybridomas TC85 and T68, respectively, concentrated 20-fold by filtration through Diaflo ultrafiltration membranes (Amicon; 10,000 molecular weight cutoff); OKT3, 125 ng of purified antibody (Ortho Diagnostic Systems, catalog no. 7032) in 95 μl of human serum. The cells were washed in PBS/0.1% FCS and resuspended in 190 μl of the same buffer containing a 1:50 dilution of FITC-conjugated goat anti-rat (Organon Teknika, catalog no. 1613–1721) for the rat primary antibodies and a 1:30 dilution of FITC-conjugated sheep anti-mouse (Organon Teknika, catalog no. 1211–1744) for OKT3. Samples were incubated on ice for 30 min and washed with PBS/0.1% FCS. Analyses were performed on an Ortho 50H Cytofluorograph (535 nm green band pass filter with an argon laser excitation of 488 nm). Histograms represent the log of fluorescence vs. cell number.

Radioiodination

Labeling was performed according to Cone and Marchalonis (1974).

Briefly, 4 × 10^7 cells were incubated at 30°C in 150 μl of PBS, 10 μl of 0.5 M sodium phosphate (pH 7.0), 1 mCi of carrier-free Na ^{125}I, 100 μg of lactoperoxidase. The reaction was initiated by the addition of 20 μl of 0.03% hydrogen peroxide, and another 20 μl was added after 6 min. The reaction was terminated 10 min after the second addition by washing with cold PBS containing 2 mM KI. For immunoprecipitation with 2C11, the cell pellets were lysed in 400 μl of lysis buffer containing 5 mM CHAPS, 0.01 M Tris (pH 8.2), 0.15 M NaCl (CHAPS buffer) supplemented with 0.1 M iodoacetamide, 5 mM phenylmethylsulfonyl fluoride, 0.5 μg of leupeptin, 1 μg of aprotonin. For immunoprecipitation with 85.4, the cell pellets were lysed in 400 μl of lysis buffer as described above, except that 0.5% Triton was used instead of CHAPS, and the Tris buffer was pH 7.3 instead of pH 8.2 (Triton buffer). After 1 hr at 0°C the nuclei and insoluble material were removed by centrifugation at 13,000 × g for 30 min.

Immunoprecipitation

Immunoprecipitations were performed using formalin-fixed Staphylococcus A (Staph A) resuspended in either CHAPS buffer with 0.4% bovine serum albumin or Triton buffer with 0.4% bovine serum albumin. Radiolabeled cell lysates were precleared by incubation at 4°C for 18 hr with Staph A that had been incubated with mouse hyperimmune sera. The Staph A was pelleted for 2 min at 13,000 × g, and the supernatants were transferred to fresh tubes. The lysates were then further precleared by three additional incubations with Staph A for 30 min, 0°C. Primary antibodies were bound to Staph A either by direct incubation with Staph A (2C11) or by incubating concentrated culture supernatant with rabbit anti-rat-coated Staph A (85.4) for 18 hr. The 2C11– or 85.4–Staph A was incubated with cell lysates for 4 hr and then collected by centrifugation. The pellets were washed in 1.5 ml of CHAPS or Triton buffer 4 times with the second and third washes containing 0.5 M NaCl. Proteins were eluted from Staph A by boiling for 2 min in SDS sample buffer with or without 2-mercaptoethanol.

Polyacrylamide Gel Electrophoresis

Two-dimensional isoelectric focusing gels were performed as described by McMillan et al. (1981). The first dimension gels contained 4% pH 3–10 ampholytes and 1.4% pH 5–7 ampholytes (LKB). The second dimension was carried out on 10% polyacrylamide gels.

Two-dimensional diagonal electrophoresis was performed exactly as described in Allison et al. (1982).

Lymphokine Release Assays

EL4, Z1, and C25 were tested for antigen reactivity using an IL-2 release assay as described by Kappler et al. (1981). Briefly, 10^5 cells in 100 μl of growth medium were incubated with 100 μl of medium containing various stimulating agents in 96-well plates. The stimulating agents were either p-aminophenyl phosphorylcholine (Sigma, catalog no. A-9278), PC*-coupled Sepharose beads (Pierce, catalog no. 20307), or 1 μl of a 1:100 dilution of 85.4 ascites fluid. After 24 hr, 100 μl of supernatant was transferred to 5 × 10^3 IL-2-dependent HT-2 cells. After an additional 24 hr, proliferation was determined by measuring the incorporation of [^{3}H]thymidine (1 μCi/well) during a 4–6 hr pulse. Maximum stimulation obtained with 20 ng of phorbol myristate acetate for all clones was between 80–100,000 cpm.

Acknowledgments

The authors thank Drs. Pamela Bjorkman, Gamal Osman, James Urban, and Dennis Zaller for critical reading of the manuscript. We also wish to thank Rochelle A. Diamond and Patrick F. Koen for expert technical assistance with the cytofluorograph and Keith D. Lewis for technical assistance with two-dimensional gels. S. M. G. and W. E. L. were supported by the Concern (II) Foundation.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received November 1, 1989; revised January 8, 1990.

References

Alarcon, B., DeVries, J., Pettey, C., Boylston, A., Yssel, H., Terhorst, C., and Spits, H. (1987). The T-cell receptor γ chain–CD3 complex: implication in the cytotoxic activity of a CD3$^-$CD4$^-$CD8$^-$ human natural killer clone. Proc. Natl. Acad. Sci. USA *84*, 3861–3865.

Allen, P., Babbitt, B. P., and Unanue, E. R. (1987). T-cell recognition of lysozyme: the biochemical basis of presentation. Immunol. Rev. *98*, 171–187.

Allison, J. P., McIntyre, B. W., and Bloch, D. (1982). Tumor-specific antigen of murine T-lymphoma defined with monoclonal antibody. J. Immunol. *129*, 2293–2300.

Amzel, L. M., and Poljak, R. J. (1979). Three-dimensional structure of immunoglobulins. Annu. Rev. Biochem. *48*, 961–998.

Arcangeli, A., Riccarda Del Bene, M., Poli, R., Ricupero, L., and Olivotto, M. (1989). Mutual contact of murine erythroleukemia cells activates depolarizing cation channels, whereas contact with extracellular substrata activates hyperpolarizing Ca^{2+}-dependent K^+ channels. J. Cell. Physiol. *139*, 1–8.

Becker, M. L. B., Near, R., Mudgett-Hunter, M., Margolies, M. N., Kubo, R. T., Kaye, J., and Hedrick, S. M. (1989). Expression of a hybrid immunoglobulin–T cell receptor protein in transgenic mice. Cell *58*, 911–921.

Borst, J., van de Griend, R. J., van Oostveen, J. W., Siew-Lan, A., Melief, C. J., Seidman, J. G., and Bolhuis, R. L. H. (1987). A T-cell receptor γ/CD3 complex found on cloned functional lymphocytes. Nature *325*, 683–688.

Brenner, M. B., Trowbridge, I. S., and Strominger, J. L. (1985). Cross-linking of human T cell receptor proteins: association between the T cell idiotype β subunit and the T3 glycoprotein heavy subunit. Cell *40*, 183–190.

Brenner, M. B., McLean, J., Pialynas, D. P., Strominger, J. L., Smith, J. A., Owen, F. L., Seidman, J. G., Ip, S., Rosen, F., and Krangel, M. S. (1986). Identification of a putative second T-cell receptor. Nature *322*, 145–149.

Chien, Y.-H., Iwashima, M., Kaplan, K. B., Elliott, J. F., and Davis, M. M. (1987). A new T-cell receptor gene located within the α locus and expressed early in T-cell differentiation. Nature *327*, 677–682.

Chiller, J. M., DeFreitas, E., Chestnut, R., Grey, H., and Skidmore, B. (1982). Signal requirement for T-lymphocyte activation. In Isolation, Characterization, and Utilization of T-Lymphocyte Clones, C. Fathman and F. Fitch, eds. (New York: Academic Press), pp. 83–100.

Clarke, C., Berenson, J., Goverman, J., Boyer, P. D., Crews, S., Siu, G., and Calame, K. (1982). An immunoglobulin promoter region is unaltered by DNA rearrangement and somatic mutation during B-cell development. Nucl. Acids Res. *10*, 7731–7749.

Clevers, H., Alarcon, B., Wileman, T., and Terhorst, C. (1988). The T-cell receptor/CD3 complex: a dynamic protein. Annu. Rev. Immunol. *6*, 629–662.

Cone, R. E., and Marchalonis, J. J. (1974). Surface proteins of thymus-derived and bone marrow-derived lymphocytes. Selective isolation of immunoglobulins and theta antigen by non-ionic detergents. Biochem. J. *140*, 345–354.

Crews, S., Griffin, J., Huang, H., Calame, K., and Hood, L. (1981). A single V_H gene segment encodes the immune response to phosphorylcholine: somatic mutation is correlated with the class of the antibody. Cell *25*, 59–66.

Davies, D., and Metzger, H. (1983). Structural basis of antibody functions. Annu. Rev. Immunol. *1*, 87–117.

Davies, D. R., Sheriff, S., and Padlan, E. A. (1988). Antibody–antigen complexes. J. Biol. Chem. *263*, 10541–10544.

Davis, M. M. (1985). Molecular genetics of the T-cell receptor β chain. Annu. Rev. Immunol. *3*, 537–560.

Davis, M. M., Berg, L. J., Lin, A., Fazekas de St. Groth, B., Devaux, B., Sagerstrom, C., Bjorkman, P. J., and Elliott, J. F. (1989). T-cell receptor recognition and selection *in vivo*. Cold Spring Harbor Symp. Quant. Biol. *LIV*, 119–128.

Desaymard, C., Giusti, A. M., and Scharff, M. D. (1984). Rat anti-T15 monoclonal antibodies with specificity for V_H and V_H–V_L epitopes. Mol. Immunol. *21*, 961–967.

Eisenberg, P., Schwarz, E., Komaromy, M., and Wall, R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J. Mol. Biol. *179*, 125–142.

Fraser, J. P., Goldsmith, M. A., and Weiss, A. (1989). Ligand-induced association between the T-cell antigen receptor and two glycoproteins. Proc. Natl. Acad. Sci. USA *86*, 7133–7137.

Gascoigne, N. R. J., Goodnow, C. C., Dudzik, K. I., Oi, U. T., and Davis, M. M. (1987). Secretion of a chimeric T-cell receptor-immunoglobulin protein. Proc. Natl. Acad. Sci. USA *84*, 2936–2940.

Goverman, J., Minard, K., Shastri, N., Hunkapiller, T., Hansburg, D., Sercarz, E., and Hood, L. (1985). Rearranged β T cell receptor genes in a helper T cell clone specific for lysozyme: no correlation between V_β and MHC restriction. Cell *40*, 859–867.

Goverman, J., Hunkapiller, T., and Hood, L. (1986). A speculative view of the multicomponent nature of T cell antigen recognition. Cell *45*, 475–484.

Hamel, P. A., Klein, M. H., Smith-Gill, S. J., and Dorrington, K. J. (1987). Relative monovalent association constant between immunoglobulin H and L chains is unrelated to their expression or antigen-binding activity. J. Immunol. *139*, 3012–3020.

Haskins, K., Hannum, C., White, J., Roehm, N., Kubo, R., Kappler, J., and Marrack, P. (1984). The antigen-specific, major histocompatibility complex-restricted receptor on T cells. VI. An antibody to a receptor allotype. J. Exp. Med. *160*, 452–471.

Herberman, R. B. (1972). Serological analysis of cell surface antigens of tumors induced by murine leukemia virus. J. Natl. Cancer Inst. *48*, 265–271.

Herrmann, S. H., and Mescher, M. F. (1986). The requirements for antigen multivalency in class I antigen recognition and triggering of primed precursor cytolytic T lymphocytes. J. Immunol. *136*, 2816–2825.

Hill, R. L., Delaney, R., Fellow, R. E., Jr., and Lebowitz, H. E. (1966). The evolutionary origins of the immunoglobulins. Proc. Natl. Acad. Sci. USA *56*, 1762–1764.

Hochstenbach, F., and Brenner, M. B. (1989). T-cell receptor δ chain can substitute for α to form a βδ heterodimer. Nature *340*, 562–565.

Hua, C., Boyer, C., Guimezanes, A., Albert, F., and Schmitt-Verhulst, A. (1986). Analysis of T-cell activation requirements with the use of alloantigens or an anti-clonotypic monoclonal antibody. J. Immunol. *136*, 1927–1936.

Hunkapiller, T., and Hood, L. (1989). Diversity of the immunoglobulin gene superfamily. Adv. Immunol. *44*, 1–62.

Kappler, J., Skidmore, B., White, J., and Marrack, P. (1981). Antigen-inducible, H-2 restricted, interleukin-2 producing T-cell hybridomas. Lack of independent antigen and H-2 recognition. J. Exp. Med. *153*, 1198–1214.

Kappler, J., Kubo, R., Haskins, K., White, J., and Marrack, P. (1983). The mouse T cell receptor: comparison of MHC-restricted receptors on two T cell hybridomas. Cell *34*, 727–737.

Kohler, H. (1975). The response to phosphorylcholine: dissecting an immune response. Transplant. Rev. *27*, 34–56.

Kourilsky, P., Chaouat, G., Rabourdin-Combe, C., and Claverie, J.-M. (1987). Working principles in the immune system implied by the "peptidic self" model. Proc. Natl. Acad. Sci. USA *84*, 3400–3404.

Kullberg, R. (1987). Stretch-activated ion channels in bacteria and animal cell membranes. Trends Neurosci. *10*, 387–388.

Kronenberg, M., Siu, G., Hood, L., and Shastri, N. (1986). The molecular genetics of the T-cell antigen receptor and T-cell antigen recognition. Annu. Rev. Immunol. *4*, 529–591.

Kuwana, Y., Asakura, Y., Utsunomiya, N., Nakanishi, M., Arata, Y., Itoh, S., Nagase, F., and Kurosawa, Y. (1987). Expression of chimeric receptor composed of immunoglobulin-derived V regions and T-cell receptor-derived C regions. Biochem. Biophys. Res. Commun. *149*, 960–968.

Lanier, L. L., Ruitenberg, J. J., Allison, J. P., and Weiss, A. (1986). Distinct epitopes on the T-cell antigen receptor of HPB-ALL tumor cells identified by monoclonal antibodies. J. Immunol. *137*, 2286–2292.

Lansman, J. B., Hallam, T. J., and Rink, T. J. (1987). Single stretch-activated ion channels in vascular endothelial cells as mechanotransducers? Nature *325*, 811–813.

Leo, O., Foo, M., Sachs, P., Samelson, L. E., and Bluestone, J. A. (1987). Identification of a monoclonal antibody specific for a murine T3 polypeptide. Proc. Natl. Acad. Sci. USA *84*, 1374–1378.

McCluskey, J., Boyd, L. F., Highet, P. F., Inman, J., and Margulies, D. H. (1988). T cell activation by purified, soluble, class I MHC molecules: requirement for polyvalency. J. Immunol. *141*, 1451–1455.

McMillan, M., Frelinger, J. A., Jones, P. P., Murphy, D. P., McDevitt, H. O., and Hood, L. (1981). Structure of murine Ia antigen. Two-dimensional electrophoretic analysis and high pressure liquid chromatography tryptic peptide maps of products of the I-A and I-E subregions and of an associated invariant peptide. J. Exp. Med. *153*, 936–950.

Meuer, S. C., Hodgdon, J. C., Hussey, R. E., Protentis, J. P., Schlossman, S. F., and Reinherz, E. L. (1983). Antigen-like effects of monoclonal antibodies directed at receptors on human T-cell clones. J. Exp. Med. *158*, 988–993.

Mizel, S. B. (1982). Interleukin I and T cell activation. Immunol. Rev. *63*, 51–72.

Morris, C. E., and Sigurdson, W. J. (1989). Stretch-inactivated ion channels coexist with stretch-activated ion channels. Science *243*, 807–809.

Mulligan, R. C., and Berg, P. (1980). Expression of a bacterial gene in mammalian cells. Science *209*, 1422–1427.

Novotny, J., Tonegawa, S., Saito, H., Kranz, D., and Eisen, H. (1986). Secondary, tertiary and quaternary structure of T-cell-specific immunoglobulin-like polypeptide chains. Proc. Natl. Acad. Sci. USA *83*, 742–746.

Oi, V. T., and Herzenberg, L. A. (1980). Immunoglobulin-producing hybrid cell lines. In Selected Methods in Cellular Immunology, B. B. Mishell and S. M., Shiigi, eds. (San Fancisco: W. H. Freeman and Company), pp. 351–372.

Parnes, J. (1989). Molecular biology and function of CD4 and CD8. Adv. Immunol. *44*, 265–311.

Rodwell, J. D., Gearhart, P. J., and Karush, F. (1983). Restriction in IgM expression. IV. Affinity analysis of monoclonal anti-phosphorylcholine antibodies. J. Immunol. *130*, 313–316.

Rojo, J., and Janeway, C. A., Jr. (1988). The biological activity of anti-T cell receptor V region monoclonal antibodies is determined by the epitope recognized. J. Immunol. *140*, 1081–1088.

Rojo, J. M., Saizawa, K., and Janeway, C. A., Jr. (1989). Physical association of CD4 and the T-cell receptor can be induced by anti-T-cell receptor antibodies. Proc. Natl. Acad. Sci. USA *86*, 3311–3315.

Rudikoff, S., Giusti, A. M., Cook, W. P., and Scharff, M. P. (1979). Single amino acid substitutions altering antigen-binding specificity. Proc. Natl. Acad. Sci. USA *79*, 1979–1983.

Saito, H., Kranz, D. M., Takagaki, Y., Hayday, A. C., Eisen, H. N., and Tonegawa, S. (1984). A third rearranged and expressed gene in a clone of cytotoxic T lymphocytes. Nature *312*, 36–40.

Saito, T., Weiss, A., Gunter, K. C., Shevach, E. M., and Germain, R. N. (1987). Cell surface T3 expression requires the presence of both α- and β-chains of the T-cell receptor. J. Immunol. *139*, 625–628.

Schneck, J., Maloy, W. L., Coligan, J. E., and Margulies, D. H. (1989). Inhibition of an allospecific T cell hybridoma by soluble class I proteins and peptides: estimation of the affinity of a T cell receptor for MHC. Cell *56*, 47–55.

Schwartz, R. H. (1984). The role of gene products of the major histocompatibility complex in T cell activation. In Fundamental Immunology, W. E. Paul, ed. (New York: Raven Press), pp. 379–438.

Segal, D. M., Padlan, E. A., Cohen, G. H., Rudikoff, S., Potter, M., and Davies, D. R. (1974). The three-dimensional structure of a phosphorylcholine-binding mouse immunoglobulin Fab and the nature of the antigen-binding site. Proc. Natl. Acad. Sci. USA *71*, 4298–4302.

Shevach, E. M., Stobo, J. D., and Green, I. (1972). Immunoglobulin and θ-bearing murine leukemias and lymphomas. J. Immunol. *108*, 1146–1151.

Southern, P. J., and Berg, P. (1982). Transformation of mammalian cells

to antibiotic resistance with a bacterial gene under control of the SV40 early region promoter. J. Mol. Appl. Genet. *1*, 327–341.

Staerz, U., Rammensee, H.-G., Benedetto, J., and Bevan, M. (1985). Characterization of a murine monoclonal antibody specific for an allotypic determinant on T-cell antigen receptor. J. Immunol. *134*, 3994–4000.

Strominger, J. L. (1989). The γδ T cell receptor and class Ib MHC-related proteins: enigmatic molecules of immune recognition. Cell *57*, 895–898.

Tycko, B., Palmer, J. P., and Sklar, J. (1989). T-cell receptor gene *trans*-rearrangements: chimeric γ–δ genes in normal lymphoid tissues. Science *245*, 1242–1246.

Watts, T. (1988). T-cell activation by preformed, long-lived Ia-peptide complexes: quantitative aspects. J. Immunol. *141*, 3708–3714.

Weiss, A., Imboden, J., Hardy, K., Manger, B., Terhorst, C., and Stobo, J. (1986). The role of the T3–antigen receptor complex in T cell activation. Annu. Rev. Immunol. *4*, 593–619.

VI.2 T-cell Co-receptor Molecules

Cell, Vol. 40, 591–597, March 1985, Copyright © 1985 by MIT

0092-8674/85/030591-07 $02.00/0

The T Cell Differentiation Antigen Leu-2/T8 Is Homologous to Immunoglobulin and T Cell Receptor Variable Regions

Vikas P. Sukhatme,*† Kurt C. Sizer,* Amy C. Vollmer,* Tim Hunkapiller,‡ and Jane R. Parnes*

* Division of Immunology
Department of Medicine
Stanford University Medical Center
Stanford, California 94305

‡ Division of Biology
California Institute of Technology
Pasadena, California 91125

Summary

Leu-2/T8 is a cell surface glycoprotein expressed by most cytotoxic and suppressor T lymphocytes. Its expression on T cells correlates best with recognition of class I major histocompatibility complex antigens, and it has been postulated to be a receptor for these proteins. We have determined the complete primary structure of Leu-2/T8 from the nucleotide sequence of its cDNA. The protein contains a classical signal peptide, two external domains, a hydrophobic transmembrane region, and a cytoplasmic tail. The N-terminal domain of the protein has striking homology to variable regions of immunoglobulins and the T cell receptor. The membrane-proximal domain appears to be a hinge-like region similar to that of immunoglobulin heavy chains. The superfamily of immunologically important surface molecules can now be extended to include Leu-2/T8.

Introduction

The Leu-2/T8 T cell differentiation antigen is a cell surface glycoprotein expressed by distinct subsets of human T lymphocytes. It is the analog and presumed homolog of mouse Lyt-2,3 (Reinherz and Schlossman, 1980; Ledbetter et al., 1981). Traditionally, these differentiation antigens have been considered markers of T cell function: cytotoxic and suppressor cells expressing Leu-2/T8 (Lyt-2,3 in mouse) and helper/inducer cells expressing the alternative antigen Leu-3/T4 (L3T4 in mouse). However, recent data from many laboratories have indicated that expression of Leu-2/T8 or Leu-3/T4 correlates not so much with function as with recognition by T cells of class I (HLA-A, B, C) or class II (HLA-DP, DQ, DR) major histocompatibility complex (MHC) antigens, respectively (Engleman et al., 1981; Krensky et al., 1982a, 1982b; Meuer et al., 1982). Leu-2/T8 is thought to be involved in the recognition by cytotoxic T cells of their targets. Monoclonal antibodies directed against Leu-2/T8 inhibit cytotoxicity by most Leu-2/T8* killer T cells (Meuer et al., 1982; Spits et al., 1982; Reinherz et al., 1983). This block is at the step of conjugate formation between the cytotoxic cell and the target cell, rather than at the later killing steps (Tsoukas et al., 1982; Landegren et al., 1982; Moretta et al., 1984). It has been suggested that the function of Leu-2/T8 (or Lyt-2,3) is to increase the avidity of, or stabilize the interaction between, the T cell and its target (Meuer et al., 1982; MacDonald et al., 1982; Moretta et al., 1984), perhaps by binding to nonpolymorphic regions of class I MHC molecules (Krensky et al., 1982a; Meuer et al., 1982; Ball and Stastny, 1982; Spits et al., 1982; Biddison et al., 1982; Engleman et al., 1983).

On peripheral T cells the Leu-2/T8 protein consists of dimers and higher multimers of a 32-34 kd glycosylated subunit, linked together by disulfide bridges (Ledbetter et al., 1981; Snow et al., 1983). On thymocytes, in addition to homodimers of this polypeptide, a 46 kd subunit is disulfide linked to the smaller chain in tetramers and higher multimers (Ledbetter et al., 1981; Snow and Terhorst, 1983). This larger subunit does not bear the T8 determinant, is unrelated to the smaller subunit by peptide mapping, and is presumably not required for the function of the molecule on mature T cells (Snow and Terhorst, 1983). Snow et al. (1984) have recently sequenced 25 N-terminal residues of Leu-2/T8 and found no significant homology to any other proteins. Determination of the complete primary structure of Leu-2/T8 would be of enormous value not only for evolutionary comparisons but also for further studies of the function of the protein in T cell recognition. To this end we have recently isolated cDNA clones that encode Leu-2/T8 (Kavathas et al., 1984). We have now determined the complete amino acid sequence of the Leu-2/T8 protein from the nucleotide sequence of a cDNA clone. We find that the protein has a classical signal peptide, two external domains, a transmembrane region, and an intracytoplasmic tail. One of the external domains has striking sequence homology to variable regions of other immunological recognition molecules. The results suggest a common evolutionary origin of Leu-2/T8, immunoglobulin, and T cell receptor genes.

Results

Nucleotide Sequence of Leu-2/T8 cDNA

The nucleotide sequence of Leu-2/T8 was determined from a single cDNA clone, pL2-M, containing a 2 kb insert, although large portions were confirmed on additional clones. Figure 1 shows a restriction map of the insert of clone pL2-M, indicating the enzyme sites used for subcloning into M13 vectors. The nucleotide sequence of this clone is presented in Figure 2. There are two long open reading frames of similar sizes at the 5′ end of this clone. One of these extends from the beginning of the clone through nucleotide 632. There are no methionines in this frame, so if it represents a translated protein, our clone is missing the initiation codon. The second open reading frame begins at nucleotide 1 and extends through nucleotide 792. This latter frame encodes Leu-2/T8, as deter-

† Present address: Department of Medicine, University of Chicago, Chicago, Illinois 60637

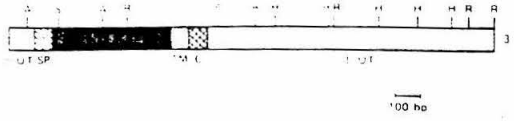


Figure 1. Restriction Map of Leu-2/T8 cDNA Clone pL2-M

A map of the 2 kb Eco RI insert of cDNA clone pL2-M is shown. The enzyme sites indicated are those used for subcloning into M13 vectors for DNA sequencing. A: Ava I. S: Sau 3A. R: Rsa I. H: Hinf I. The artificial Eco RI sites at the 5' and 3' ends are not labeled. The domains of the protein encoded by various segments of the cDNA are demarcated. Open bar: 5' and 3' untranslated (UT) regions. Hatched: signal peptide (SP). Shaded: external protein domains. Speckled: transmembrane region (TM). Crosshatched: cytoplasmic tail (C).

mined by comparison with the N-terminal protein sequence (Snow et al., 1984). The first ATG in this frame is at nucleotide 88. We believe this to be the initiation codon. Accordingly, clone pL2-M contains 115 bp of 5' untranslated region (not all shown), 705 bp of protein coding sequence (235 amino acids), and 1182 bp of 3' untranslated region. Neither this clone nor another extending 85 bp more 3' contains a poly(A) tail. The mRNA represented by these clones is approximately 2.5 kb in length (Kavathas et al., 1984). Our clones span 2.1 kb, so we are missing approximately 400 bp. We presume that most of this is 3' untranslated region and poly(A) tail.

Figure 2. Nucleotide Sequence and Deduced Amino Acid Sequence of the Insert of Leu-2/T8 cDNA Clone pL2-M

Horizontal arrows denote the beginning of the predicted protein domains. Open circles are placed above the two cysteines in the V-like domain that are conserved in immunoglobulin and T cell receptor V regions. A closed circle is above the potential site for N-linked glycosylation. A horizontal bar is over the 2° amino acid sequence homologous to the mouse IgA hinge. Numbers in the left and right margins, respectively, refer to the first and last nucleotides or amino acids in each line. Abbreviations: L, leader (signal peptide); V, variable region-like domain; H, hinge; TM, transmembrane region; C, cytoplasmic tail; 3'UT, 3' untranslated region.

Figure 3. Comparison of the Amino Acid Sequence of the N-Terminal Domain of Leu-2/T8 with Immunoglobulin and T Cell Receptor Variable Region Sequences

The Leu-2/T8 sequence is the first 96 amino acids of the mature protein. Similarities to Leu-2/T8 are denoted by a dot (.); gaps in the alignments are indicated by blank spaces. Immunoglobulin lambda, kappa, and heavy chain V regions are denoted V_λ, V_κ, and V_H. T cell receptor α and β chain V regions are denoted V_α and V_β. The sequence alignments were derived by comparing all members of a given class of V regions (e.g. kappa, lambda) against one another and Leu-2/T8 simultaneously to achieve the best alignment for a group of sequences, with attention paid to alignment of certain key conserved residues. All of the immunoglobulin V sequences in the Dayhoff bank were utilized, as were all published and some unpublished T cell receptor sequences (L. E. Hood, personal communication). The sequences were from the following sources: human Ha V_λ-I (Shinoda et al., 1970); human Mcg V_λ-II (Fett and Deutsch, 1974); mouse U61 V_κ (Vrana et al., 1979); human Len V_κ (Schneider and Hilschmann, 1975); mouse V_H (Capra and Nisonoff, 1979); rat IR2 V_H (Heilman et al., 1982); mouse TT11 V_β (Chien et al., 1984); human TY35 V_α (Yanagi et al., 1984).

Predicted Amino Acid Sequence of Leu-2/T8

The predicted amino acid sequence of Leu-2/T8 is shown in Figure 2. The protein begins with a 21 amino acid hydrophobic leader sequence, presumably representing a signal peptide that is cleaved off as the protein passes into the endoplasmic reticulum. The indicated start of the mature processed protein is based upon comparison with the published N-terminal protein sequence (Snow et al., 1984). Our N-terminal sequence differs from the latter in only two places. We predict a glutamine at residue 2, while the protein sequence indicates glutamic acid at this position. As the cDNA was sequenced multiple times on both strands, we believe that either this residue became deamidated during the protein purification or there was an error made by reverse transcriptase during the cDNA cloning. Snow et al. (1984) placed a gap of one amino acid between the cysteine at residue 22 and the two leucines at what were considered residues 24 and 25. We predict two amino acids in this region: glutamine at residue 23 and valine at 24. This discrepancy is likely to be a protein sequencing error, since it is at the very end of the determined sequence. However, it is possible that either or both of these discrepancies could be the result of polymorphism.

The mature Leu-2/T8 protein consists of 214 amino acids. A 24 amino acid hydrophobic segment begins at position 162. We believe this to represent a transmembrane region. The following 29 amino acid C-terminal segment (residues 186–214) is highly charged (ten basic residues and only one acidic residue) and is presumably an intracytoplasmic tail. Our sequence predicts an unglycosylated mature protein chain of 23,554 daltons as compared to the described 32,000–34,000 dalton glycosylated subunit size (Ledbetter et al., 1981; Snow and Terhorst, 1983). However, when Snow et al. (1984) treated the isolated protein with trifluoromethane sulfonic acid (TFMS) to remove both N-linked and O-linked carbohydrate, the size was reduced only 1.5-2 kd. The reasons for this discrepancy are unclear, but most likely relate to incomplete cleavage of carbohydrates from the protein with TFMS. Although our sequence predicts one potential site for N-linked glycosylation (Arg-X-Thr or Arg-X-Ser), at position 28, it may not be used, since experiments with endoglycosidase F suggest that there is no N-linked glycosylation (Snow and Terhorst, 1983).

Leu-2/T8 Is Homologous to Immunoglobulin and T Cell Receptor Variable Regions

We searched the Dayhoff protein sequence data bank for homologies of the predicted Leu-2/T8 protein sequence to other proteins. The greatest amount of homology found was to immunoglobulin light chain variable (V) regions, with less but still significant homology to immunoglobulin heavy chain V regions (V_H) and V regions of the α and β chains of the T cell receptor. Comparisons of the Leu-2/T8 protein sequence with examples of immunoglobulin and T cell receptor V region sequences are illustrated in Figure 3. The segment of Leu-2/T8 that is homologous to a V region is at the N terminus and extends to amino acid 96. For the V regions the homology to Leu-2/T8 stops just before the D (diversity) segment (if one is present) or the J (joining) segment. With the alignment shown the homology is on the order of 30%–35% to lambda and kappa V regions, 20%–22% to V_H, and 24% to T cell receptor α and β chain V regions. The homology increases substantially (e.g. up to 56% for lambda or 58% for kappa) if one includes conservative amino acid substitutions. As shown in Figure 3, there are seven amino acids that are conserved in all of these sequences. These include the two cysteine residues involved in the classical intrachain disulfide loop of V regions (positions 22 and 94 for Leu-2/T8) and the tryptophan at position 35, which is thought to be important for the proper folding of immunoglobulin domains. Leu-2/T8 contains most of the residues of light chain V regions that are used for association with heavy chains (Poljak et al., 1975), as well as many of the V region contact residues in light chain dimers (Davies et al., 1975). We therefore predict that the two V-like domains in Leu-2/T8 dimers are associated with one another (noncovalently), and probably form a single binding site for a presumed ligand. As might be expected, the homology of Leu-2/T8 to V regions is predominantly to framework rather than hypervariable regions. The level of homology does not vary substantially across mammalian species.

In contrast to the findings for the N-terminal external domain, the 65 amino acid membrane-proximal domain of Leu-2/T8 bears no strong resemblance to constant region domains of immunoglobulins or the T cell receptor and is not homologous to other known proteins. However, a stretch of 21 amino acids in the center of this domain (residues 120–140) is homologous to the hinge region of

Figure 4. Homology of Leu-2/T8 Hinge to Mouse IgA Hinge

Amino acids 120–140 of Leu-2/T8 are shown. Amino acids 220–240 of the mouse α heavy chain (CH_1) are below. The entire CH_1 hinge spans amino acids 218–240. Homologies are denoted by an asterisk (*), and gaps by blank spaces. The CH_1 hinge sequence is from Auffray et al. (1981).

the mouse IgA heavy chain (Figure 4). Although we doubt an evolutionary significance of this homology, evidence discussed below suggests that the domain containing this sequence is a hinge. The remaining protein domains appear unique to Leu-2/T8.

Structural Predictions

We have analyzed the Leu-2/T8 sequence using a formula similar to that of Kyte and Doolittle (1982) for determining the degree of hydrophobicity or hydrophilicity of segments of a protein. Such analyses allow one to predict which portions of a protein are exposed, interior, or embedded in a membrane. A hydrophobicity plot of the Leu-2/T8 sequence is shown in Figure 5. The putative signal peptide (residues −21 to −1) and transmembrane domain (residues 162 to 185) are both demonstrated to be highly hydrophobic, while the presumed intracytoplasmic tail (residues 186 to 214) is very hydrophilic. The V-like N-terminal external domain has alternating stretches of hydrophobic and hydrophilic residues as seen in globular proteins (Kyte and Doolittle, 1982). Furthermore, the plot for this domain is virtually superimposable on those for kappa, lambda, and heavy chain V regions, except for an extra hydrophilic segment in Leu-2/T8 between amino acids 6 and 15 (data not shown). These results suggest a great deal of conservation of structure among these protein domains. We further analyzed the Leu-2/T8 protein sequence according to the parameters of Chou and Fasman (1978) and find that the predicted β-sheet structure of the N-terminal domain is the same as that for light chain V regions. In contrast, the membrane-proximal external domain does not fit into any structure. Therefore this domain, which was shown to contain a sequence homologous to the mouse IgA hinge, is itself predicted to be a hinge region.

The Leu-2/T8 Gene Does Not Rearrange in Cells That Express the Protein

Variable regions of immunoglobulins and the T cell receptor are not expressed as functional receptor molecules until their genes rearrange during the development of B cells or T cells, respectively, and become juxtaposed to either a J or D and J segments just upstream from a constant region gene. Since these proteins are homologous to Leu-2/T8, it was important to determine whether the Leu-2/T8 gene also rearranges in T cells that express this protein. We have therefore compared Southern blots of DNA from human placenta (which does not express Leu-2/T8 protein or mRNA) and from a Leu-2/T8-expressing thymoma cell line, JM, hybridized to the insert of cDNA clone pL2-M. As illustrated in Figure 6, no difference could be detected between the two DNAs with any of the four enzymes used (Bam HI, Eco RI, Apa I, and Eco RV). There is also no difference seen with Hind III (data not shown). Since the probe contains the entire coding sequence as well as most of the noncoding portions, we conclude that no major rearrangements are required for expression of this gene. Although we cannot exclude very small rearrangements with either little or no deletion of DNA, immunoglobulin and T cell receptor type rearrangements would have been easily detectable.

The data shown in Figure 6 also suggest that Leu-2/T8 is encoded by a single gene. The Southern blots show one to three hybridizing bands, depending upon the enzyme. Differential hybridization to 5′ and 3′ probes indicates that when multiple bands are present, they represent portions of a single gene, and all of the hybridizing fragments can be accounted for on a single genomic clone containing a 14 kb insert (data not shown). Of course, it cannot be excluded that the entire Leu-2/T8 gene and large amounts of flanking sequence are repeated in the genome with little or no change. Using low stringency wash conditions (37°C, 1× SSC) we have detected one or two extra bands on genomic blots with DNA digested with Bam HI, but not reproducibly with other enzymes. We are currently determining whether these are truly distantly cross-hybridizing genes or aberrant digestion products of the Leu-2/T8 gene with Bam HI. It is clear, however, that Leu-2/T8 is not a member of a large multigene family.

Discussion

The structure of Leu-2/T8 is typical of a cell surface protein. It contains a hydrophobic signal peptide, two predicted external protein domains, a transmembrane hydrophobic domain, and an extremely basic intracytoplasmic tail. The protein has a total of nine cysteine residues: three in the N-terminal V-like domain and two each in the membrane-proximal (hinge), the transmembrane, and the cytoplasmic domains. Biochemical data indicate that the three in the V-like domain are not used for interchain disulfide bridging (Snow and Terhorst, 1983). Since our alignment with V regions shows conservation of the cysteines at positions 22 and 94 of Leu-2/T8, we predict that these two cysteines are used for an intrachain disulfide bridge as in immunoglobulins and the T cell receptor. The function of the cysteine at position 33 is unknown. Subunit joining could occur between membrane-proximal domains, within the membrane, or within the cytoplasm. Several of these six cysteine residues may be used for interchain bridging, since the molecule often forms large multimers.

Perhaps the most interesting finding concerning the structure of Leu-2/T8 is the homology of the N-terminal external domain to variable regions. Many members of the "superfamily" (Jensenius and Williams, 1982; Williams, 1984) of immunologically important surface proteins are much more homologous to constant regions than to variable regions (e.g. class I and class II MHC antigens, β_2-microglobulin). In contrast to these proteins, Leu-2/T8 is

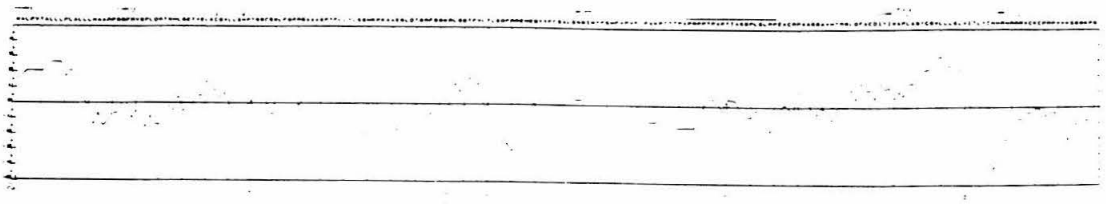


Figure 5. Hydrophobicity Plot of Leu-2/T8 Protein Sequence

Data are plotted according to Kyte and Doolittle (1982). The plotted values are averages over groups of seven contiguous residues. Positive values indicate hydrophobic, and negative values indicate hydrophilic, regions of the protein. The amino acid sequence is listed above the plot and numbered below. The domains of the protein are labeled as in the legend to Figure 2.

thought to be a receptor molecule, and so it is not surprising that it would borrow or share a structure that the immune system has used to such advantage in antigen recognition. It should be stressed, however, that there is no evidence for variability of this "V-like" domain. Two other members of the superfamily (besides immunoglobulins and T cell receptors) are thought to be related to variable regions: the mouse T cell marker Thy-1 (Williams and Gagnon, 1982) and the poly(Ig) receptor (Mostov et al., 1984). The homology of Leu-2/T8 to these molecules is less striking, as is their homology to immunoglobulin V regions.

Our sequence analyses indicate that Leu-2/T8 is about equidistant between its two closest relatives: kappa and lambda V regions. If one were to construct a genealogical tree of the immunoglobulin-related superfamily, Leu-2/T8 would split off after the divergence of heavy and light chains, and presumably just before the split between kappa and lambda. Chromosomal mapping studies of Leu-2/T8 are intriguing in this regard. Lyt-2.3 has been known for many years to be tightly linked to the kappa locus on mouse chromosome 6 (Itakura et al., 1972; Gottlieb, 1974; Claflin et al., 1978; Gibson et al., 1978). We have recently shown that in the human system the Leu-2/T8 gene also maps extremely close to the kappa locus on chromosome 2 (Sukhatme et al., 1985). This close linkage between the genes encoding kappa and Leu-2/T8, which is maintained across at least two species, lends credence to the idea of a common evolutionary origin of these proteins.

Although the N-terminal domain of Leu-2/T8 is remarkably similar to V regions, it differs from them in several important aspects. First, the gene encoding Leu-2/T8 shows no evidence of rearrangement in cells that express the protein. Second, the Leu-2/T8 gene has no close relatives in the human genome, in contrast to most, though not all, V genes. Finally, the nucleotide sequence encoding Leu-2/T8 is strikingly distinct from the sequences encoding V regions in two aspects. As is true for most protein-coding sequences (Fickett, 1982), the G+C content of the region encoding the V-like domain (or the entire coding region) of Leu-2/T8 is high (66%). In contrast, the G+C content of immunoglobulin V regions is much closer to random (T. Hunkapiller and L. E. Hood, unpublished results). Furthermore, we have analyzed the Leu-2/T8 cDNA sequence with an algorithm similar to that of Fickett (1982), which can distinguish protein coding sequences from noncoding DNA. This test for coding sequences is based upon the nonrandom use of codons, resulting in repetition of particular nucleotides with a periodicity of 3 in coding, but not noncoding, DNA. The sequence encoding Leu-2/T8 is clearly predicted as protein-coding, as is that of the V-like domain by itself. In contrast, immunoglobulin V region sequences (but not constant region sequences) are an exception to the rule in that they are most often predicted as noncoding (Fickett, 1982; T. Hunkapiller and L. E. Hood, unpublished results). These results imply that there is less constraint on the evolution of V genes than

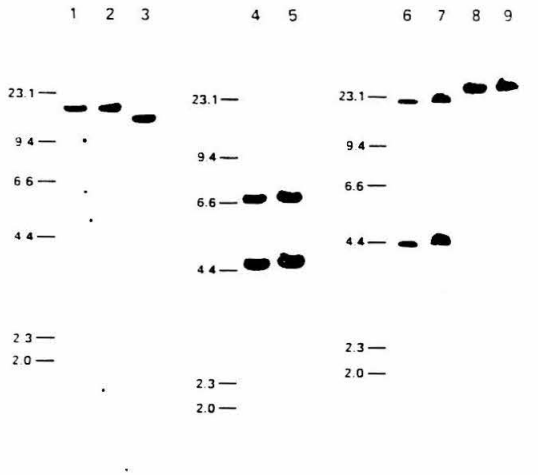


Figure 6. Southern Blot Analysis of Leu-2/T8 Gene

Genomic DNA (8 μg) from a human thymoma cell line (JM) that expresses Leu-2/T8 (lanes 1, 3, 4, 6, 8) and from human placenta, which does not express Leu-2/T8 (lanes 2, 5, 7, 9), were digested with Bam HI (lanes 1 and 2), Hind III (lane 3), Eco RI (lanes 4 and 5), Apa I (lanes 6 and 7), or Eco RV (lanes 8 and 9) and electrophoresed on a 0.8% agarose gel. Southern blots of the gels were hybridized to the insert of cDNA clone pL2-M labeled with ^{32}P by nick-translation. The migration distances of Hind III fragments of λ phage DNA are indicated in the left margins in kilobase pairs.

on other genes and that, despite its homology to V regions, Leu-2/T8 is under much stronger selective pressure to maintain its sequence.

Experimental Procedures

Isolation of cDNA Clone pL2-M

The inserts of two Leu-2/T8 cDNA clones isolated previously by subtractive hybridization (Kavathas et al., 1984) were labeled with ^{32}P by nick translation and used as probes to rescreen a previously described λgt10 cDNA library. This library was constructed from RNA of L cells that had been transfected with total human genomic DNA and selected for expression and amplification of the Leu-2/T8 gene (Kavathas et al., 1984). Positive clones were screened for insert size. The longest insert was 2 kb. This clone was grown up, and the Eco RI insert was isolated and subcloned into pBR322. The resulting plasmid is referred to as cDNA clone pL2-M.

DNA Sequencing

The Eco RI insert of cDNA clone pL2-M was digested with Rsa I and inserted into Eco RI plus Sma I digested M13 vectors (Messing et al., 1981) mp10 and mp11, or into Sma I digested mp11. The insert was also digested with Hinf I, the 5′ overhangs filled-in with the Klenow fragment of DNA polymerase I (Bethesda Research Laboratories), and the fragments subcloned into Sma I digested mp11. The largest Hinf I digestion product (the 5′ 1.1 kb Eco RI–Hinf I fragment) was isolated from the pL2-M insert, digested with Ava I and Sau 3A, filled-in with Klenow polymerase, and inserted into Sma I digested mp11. Finally, the entire Eco RI inserts of clone pL2-M and four independent, overlapping cDNA clones were cloned into the Eco RI site of mp11. Except for the first 48 bp of 5′ untranslated sequence, which were only sequenced on one strand, the nucleotide sequences of all these subclones were determined on both strands multiple times by the dideoxynucleoside triphosphate chain termination method of Sanger et al. (1977). The sequence shown was all derived from clone pL2-M, although the partial sequences of other independent cDNA clones agreed entirely except for a single base pair in the 3′ untranslated region (a run of six As beginning at nucleotide 1895 in pL2-M and seven As in clone pL2-4). The M13 vectors were obtained from Amersham.

Southern Blots and Hybridization

Genomic DNA was isolated from the T cell lymphoma line JM, by the citric acid procedure (Hieter et al., 1981). Human placental DNA was a generous gift from Dr. Philip Leder (Harvard Medical School). DNA was digested with restriction enzymes (New England Biolabs) for 3 hr and electrophoresed on 0.8% agarose gels. Gels were blotted onto nitrocellulose by the procedure of Southern (1975). Blots were hybridized for 16 hr in 40% formamide, 20 mM Tris-HCl (pH 7.6), 4× SSC, 1× Denhardt's solution, 0.1% SDS, and 30 μg/ml denatured herring sperm DNA, to the insert from clone pL2-M labeled with ^{32}P by nick translation to a specific activity of 200–400 cpm per picogram. The filters were washed three times at room temperature for 20 min each with 2× SSC, 0.05% SDS, and twice at 52°C for 1 hr each with 0.1× SSC, 0.05% SDS. Exposures with XAR-5 film (Kodak) were overnight at −70°C with intensifying screens.

Acknowledgments

We thank Dr. Philip Leder for the human placental DNA, Dr. Rose Zamoyska for critical reading of the manuscript, Dr. Leroy Hood for helpful discussions and for use of his computer facilities, and Diane O'Neill for preparation of the manuscript. This work was supported by NIH grant AI 19512 to J. R. P. V. P. S. was a National Institutes of Health postdoctoral fellow. J. R. P. is the recipient of a John A. and George L. Hartford faculty fellowship award.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received December 18, 1984; revised January 10, 1985

References

Auffray, C., Nageotte, R., Sikorav, J.-L., Heidmann, O., and Rougeon, F. (1981). Mouse immunoglobulin A: nucleotide sequence of the structural gene for the α heavy chain derived from cloned cDNAs. Gene *13*, 365–374.

Ball, E. J., and Stastny, P. (1982). Cell-mediated cytotoxicity against HLA-D region products expressed in monocytes and B lymphocytes. IV. Characterization of effector cells using monoclonal antibodies against human T cell subsets. Immunogenetics *16*, 157–169.

Biddison, W. E., Rao, P. E., Talle, M. A., Goldstein, G., and Shaw, S. (1982). Possible involvement of the T4 molecule in T cell recognition of class II HLA antigens: evidence from studies of cytotoxic T lymphocytes specific for SB antigens. J. Exp. Med. *156*, 1065–1083.

Capra, J. D., and Nisonoff, A. (1979). Structural studies on induced antibodies with defined idiotypic specificities. VII. The complete amino acid sequence of the heavy chain variable region of anti-*p*-azophenylarsonate antibodies from A/J mice bearing a cross-reactive idiotype. J. Immunol. *123*, 279–284.

Chien, Y., Becker, D. M., Lindsten, T., Okamura, M., Cohen, D. I., and Davis, M. M. (1984). A third type of murine T-cell receptor gene. Nature *312*, 31–35.

Chou, P. Y., and Fasman, G. D. (1978). Empirical predictions of protein conformation. Ann. Rev. Biochem. *47*, 251–276.

Claflin, J. L., Taylor, B. A., Cherry, M., and Cubberly, M. (1978). Linkage in mice of genes controlling an immunoglobulin kappa-chain marker and the surface alloantigen Ly-3 on T lymphocytes. Immunogenetics *6*, 379–387.

Davies, D. R., Padlan, E. A., and Segal, D. M. (1975). Three-dimensional structure of immunoglobulins. Ann. Rev. Biochem. *44*, 639–667.

Engleman, E. G., Benike, C. J., Grumet, F. C., and Evans, R. L. (1981). Activation of human T lymphocyte subsets: helper and suppressor/cytotoxic T cells recognize and respond to distinct histocompatibility antigens. J. Immunol. *127*, 2124–2129.

Engleman, E. G., Benike, C. J., Metzler, C., Gatenby, P. A., and Evans, R. L. (1983). Blocking of human T lymphocyte functions by anti-Leu-2 and anti-Leu-3 antibodies. J. Immunol. *130*, 2623–2628.

Fett, J. W., and Deutsch, H. F. (1974). Primary structure of the Mcg λ chain. Biochemistry *13*, 4102–4114.

Fickett, J. W. (1982). Recognition of protein coding regions in DNA sequences. Nucl. Acids Res. *10*, 5303–5318.

Gibson, D. M., Taylor, B. A., and Cherry, M. (1978). Evidence for close linkage of a mouse light chain marker with the Ly-2.3 locus. J. Immunol. *121*, 1585–1590.

Gottlieb, P. (1974). Genetic correlation of a mouse light chain variable region marker with a thymocyte surface antigen. J. Exp. Med. *140*, 1432–1437.

Hellman, L., Petterson, U., Engstrom, A., Karlsson, T., and Bennich, H. (1982). Structure and evolution of the heavy chain from rat immunoglobulin E. Nucl. Acids Res. *10*, 6041–6049.

Hieter, P. A., Hollis, G. F., Korsmeyer, S. J., Waldmann, T. A., and Leder, P. (1981). Clustered arrangement of immunoglobulin λ constant region genes in man. Nature *294*, 536–540.

Itakura, K., Hutton, J. J., Boyse, E. A., and Old, L. J. (1972). Genetic linkage relationships of loci specifying differentiation alloantigens in the mouse. Transplantation *13*, 239–243.

Jensenius, J. C., and Williams, A. F. (1982). The T lymphocyte antigen receptor–paradigm lost. Nature *300*, 583–588.

Kavathas, P., Sukhatme, V. P., Herzenberg, L. A., and Parnes, J. R. (1984). Isolation of the gene coding for the human T lymphocyte antigen Leu-2 (T8) by gene transfer and cDNA subtraction. Proc. Natl. Acad. Sci. USA *81*, 7688–7692.

Krensky, A. M., Clayberger, C., Reiss, C. S., Strominger, J. L., and Burakoff, S. J. (1982a). Specificity of OKT4 cytotoxic T lymphocyte clones. J. Immunol. *129*, 2001–2003.

Krensky, A. M., Reiss, C. S., Mier, J. W., Strominger, J. L., and Burakoff,

S. J. (1982b). Long-term human cytolytic T-cell lines allospecific for HLA-DR6 antigen are OKT4+. Proc. Natl. Acad. Sci. USA *79*, 2365–2369.

Kyte, J., and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. *157*, 105–132.

Landegren, U., Ramstedt, U., Axberg, I., Ullberg, M., Jondal, M., and Wigzell, H. (1982). Selective inhibition of human T cell cytotoxicity at levels of target recognition or initiation of lysis by monoclonal OKT3 and Leu-2a antibodies. J. Exp. Med. *155*, 1579–1584.

Ledbetter, J. A., Evans, R. L., Lipinski, M., Cunningham-Rundles, C., Good, R. A., and Herzenberg, L. A. (1981). Evolutionary conservation of surface molecules that distinguish T lymphocyte helper/inducer and cytotoxic/suppressor subpopulations in mouse and man. J. Exp. Med. *153*, 310–323.

MacDonald, H. R., Glasebrook, A. L., Bron, C., Kelso, A., and Cerottini, J. C. (1982). Clonal heterogeneity in the functional requirement for Lyt-2/3 molecules on cytolytic T lymphocytes (CTL): possible implications for the affinity of CTL antigen receptors. Immunol. Rev. *68*, 89–115.

Messing, J., Crea, R., and Seeburg, P. H. (1981). A system for shotgun DNA sequencing. Nucl. Acids Res. *9*, 309–321.

Meuer, S. C., Hussey, R. E., Hodgdon, J. C., Hercend, T., Schlossman, S. F., and Reinherz, E. L. (1982). Surface structures involved in target recognition by human cytotoxic T lymphocytes. Science *218*, 471–473.

Moretta, A., Pantaleo, G., Mingari, M. C., Moretta, L., and Cerottini, J. C. (1984). Clonal heterogeneity in the requirement for T3, T4, and T8 molecules in human cytolytic T lymphocyte function. J. Exp. Med. *159*, 921–934.

Mostov, K. E., Friedlander, M., and Blobel, G. (1984). The receptor for transepithelial transport of IgA and IgM contains multiple immunoglobulin-like domains. Nature *308*, 37–43.

Poljak, R. J., Amzel, L. M., Chen, B. L., Phizackerley, R. P., and Saul, F. (1975). Structural basis for the association of heavy and light chains and the relation of subgroups to the conformation of the active site of immunoglobulins. Immunogenetics *2*, 393–394.

Reinherz, E. L., and Schlossman, S. F. (1980). The differentiation and function of human T lymphocytes. Cell *19*, 821–827.

Reinherz, E. L., Meuer, S. C., and Schlossman, S. F. (1983). The human T cell receptor: analysis with cytotoxic T cell clones. Immunological Rev. *74*, 83–112.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA *74*, 5463–5467.

Schneider, M., and Hilschmann, N. (1975). The primary structure of a monoclonic immunoglobulin-L-chain of subgroup IV of the kappa type (Bence-Jones protein Len). Hoppe-Seylers Z. Physiol. Chem. *356*, 507–557.

Shinoda, T., Titani, K., and Putman, F. W. (1970). Amino acid sequence of human κ chains. II. Chymotryptic peptides and sequence of protein Ha. J. Biol. Chem. *245*, 4475–4487.

Snow, P., Spits, H., deVries, J., and Ternorst, C. (1983). Comparison of target antigens of monoclonal reagents OKT5, OKT8 and Leu-2a, which inhibit effector function of human cytotoxic T lymphocytes. Hybridoma *2*, 187–199.

Snow, P. M., and Ternorst, C. (1983). The T8 antigen is a multimeric complex of two distinct subunits on human thymocytes but consists of homomultimeric forms on peripheral blood T lymphocytes. J. Biol. Chem. *258*, 14675–14681.

Snow, P. M., Kelzer, G., Coligan, J. E., and Terhorst, C. (1984). Purification and N-terminal amino acid sequence of the human T cell surface antigen T8. J. Immunol. *133*, 2058–2066.

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. *98*, 503–517.

Spits, H., Ijssel, H., Ternorst, C., and deVries, J. E. (1982). Establishment of human T lymphocyte clones highly cytotoxic for an EBV-transformed B cell line in serum-free medium: isolation of clones that differ in phenotype and specificity. J. Immunol. *128*, 95–99.

Sukhatme, V. P., Vollmer, A. C., Erikson, J., Isobe, M., Croce, C., and Parnes, J. R. (1985). The gene for the human T cell differentiation antigen Leu-2/T8 is closely linked to the κ light chain locus on chromosome 2. J. Exp. Med. *161*, 429–434.

Tsoukas, C. D., Carson, D. A., Fong, S., and Vaughan, J. H. (1982). Molecular interactions in human T cell-mediated cytotoxicity to EBV II. Monoclonal antibody OKT3 inhibits a post-killer target recognition/adhesion step. J. Immunol. *129*, 1421–1425.

Vrana, M., Rudikoff, S., and Potter, M. (1979). The structural basis of a hapten-inhibitable κ-chain idiotype. J. Immunol. *122*, 1905–1910.

Williams, A. F. (1984). The immunoglobulin superfamily takes shape. Nature *308*, 12–13.

Williams, A. F., and Gagnon, J. (1982). Neuronal cell Thy-1 glycoprotein: homology with immunoglobulin. Science *216*, 696–703.

Yanagi, Y., Yoshikai, Y., Leggett, K., Clark, S., Aleksander, I., and Mak, T. (1984). A human T cell-specific cDNA clone encodes a protein having extensive homology to immunoglobulin chains. Nature *308*, 145–149.

Reprint Series
31 October 1986, Volume 234, pp. 610–614

SCIENCE

Isolation and Sequence of L3T4 Complementary DNA Clones: Expression in T Cells and Brain

BÉATRICE TOURVIEILLE, SCOTT D. GORMAN, ELIZABETH H. FIELD, TIM HUNKAPILLER, AND JANE R. PARNES

Isolation and Sequence of L3T4 Complementary DNA Clones: Expression in T Cells and Brain

BÉATRICE TOURVIEILLE, SCOTT D. GORMAN, ELIZABETH H. FIELD, TIM HUNKAPILLER, JANE R. PARNES

T lymphocytes express on their surface not only a specific receptor for antigen and major histocompatibility complex proteins, but also a number of additional glycoproteins that are thought to play accessory roles in the processes of recognition and signal transduction. L3T4 is one such T-cell surface protein that is expressed on most mouse thymocytes and on mature mouse T cells that recognize class II (Ia) major histocompatibility complex proteins. Such cells are predominantly of the helper/inducer phenotype. In this study, complementary DNA clones encoding L3T4 were isolated and sequenced. The predicted protein sequence shows that L3T4 is a member of the immunoglobulin gene superfamily. It is encoded by a single gene that does not require rearrangement prior to expression. Although the protein has not previously been demonstrated on nonhematopoietic cells, two messenger RNA species specific for L3T4 are found in brain. The minor species comigrates with the L3T4 transcript in T cells, whereas the major species is 1 kilobase smaller.

MATURE MOUSE T LYMPHOCYTES can be divided into two subsets by their expression of the alternative T-cell differentiation antigens L3T4 (CD4 in humans) and Lyt-2 (CD8 in humans). The L3T4 subset consists predominantly of helper/inducer T cells and correlates best with recognition by T cells of class II (Ia) major histocompatibility complex (MHC) molecules (*1–3*). The Lyt-2 subset is made up primarily of cytotoxic and suppressor T cells and correlates best with recognition of class I (H-2K, D, or L) MHC molecules (*3, 4*). Monoclonal antibodies specific for L3T4 or Lyt-2 inhibit the functional activity (cytotoxicity, proliferation, lymphokine release) of T cells that bear these proteins (*1–6*). It has been postulated that L3T4 and Lyt-2 play a role in increasing the avidity of the interaction between T cells and antigen-presenting cells or target cells, perhaps by binding to nonpolymorphic regions of class II and class I MHC proteins, respectively (*2, 3, 6, 7*). It has alternatively been postulated that monoclonal antibodies specific for L3T4 (or CD4) inhibit function by directly transmitting a negative signal to the T cell (*8*). Although Lyt-2 and CD8 are normally known to be expressed only on particular subsets of thymocytes, T cells, and natural killer cells, CD4 has also been shown to be expressed by normal cells of the monocyte/macrophage and Langerhans lineages in both humans and rats (*9*). However, none of these proteins has been reported to be expressed on normal nonhematopoietic cells.

The genes encoding human CD8, CD4, and mouse Lyt-2 have been recently cloned, and their predicted amino acid sequences have revealed that they are evolutionarily related to immunoglobulin (Ig) variable (V) regions (*10–14*). We have now cloned the complementary DNA (cDNA) encoding mouse L3T4 and show that it too is a member of the Ig gene superfamily. We further show that the gene is expressed not only in T-lineage cells but also in brain and that the size of the mRNA in brain is different from that in T cells.

To isolate mouse L3T4 cDNA clones we screened a C57BL/Ka mouse thymocyte cDNA library with a full-length human CD4 cDNA clone (*12*) used as probe. Two mouse clones that hybridized to the human clone were isolated. The nucleotide sequence of the one (pcL3T4-C7) with the longer insert (1.3 kb) was determined (Fig. 1). Because this clone did not contain the 5′ end, it was used as a probe to isolate from the same library an additional cDNA clone (pcL3T4-14) that extended farther in the 5′ direction. The nucleotide sequence of the 5′ untranslated region, the leader, and the first ten amino acids of the mature protein were therefore determined from pcL3T4-14 (Fig. 1). The nucleotide sequence shown in Fig. 1 predicts a mature protein of 435 amino acids (predicted molecular size 48,853 daltons), with 372 amino acids external to the cell, a 25–amino acid hydrophobic transmembrane region, and a 38–amino acid highly basic cytoplasmic domain. The mature protein sequence is preceded by a 22–amino acid hydrophobic leader or signal peptide as is typically found at the NH_2-terminus of cell surface and secreted proteins. The point of cleavage of this leader was predicted by comparison with other published leader sequences.

As expected, the nucleotide sequence of the L3T4 cDNA was homologous to that of the human CD4 clone with which it was selected, and the encoded protein was also closely related (Fig. 2A). The most highly conserved region was the cytoplasmic domain (79% at the amino acid level), which may play a role in signal transduction. In contrast, the external portions of L3T4 and CD4 contained only 55% identical residues. This latter finding is similar to our previous results comparing the mouse (Lyt-2) and human (CD8) sequences of the alternative T-cell differentiation marker (*13*). The mouse L3T4 protein has four predicted *N*-linked glycosylation sites (Asn-X-Thr or Asn-X-Ser) at residues 165, 276, 301, and 370, as compared to only two in the human CD4 (*12*).

We and others observed previously that CD8 and Lyt-2 have NH_2-terminal external domains that are homologous to the Ig light chain V regions (*11, 13, 14*). A similar relation has been found for human CD4 (*12*). We therefore searched a series of data banks with the L3T4 sequence to see whether similar or additional homologies could be found. These computer comparisons indicated that L3T4 is also a member of the Ig gene superfamily. The NH_2-terminal domain of the mature protein (90 to 101 amino acids, depending on where one arbitrarily sets the border) is homologous to Ig V regions, with the greatest similarity being to light chain V regions, especially κ (up to 35%) (Fig. 2B). This domain of L3T4 has the two cysteines (residues 20 and 90) that form the characteristic disulfide loop of Ig-like homology units, as well as the structurally important tryptophan (residue 32) that is always found 12 to 15 residues downstream from the first cysteine of the disulfide

B. Tourvieille, S. D. Gorman, E. H. Field, J. R. Parnes, Department of Medicine, Division of Immunology, Stanford University Medical Center, Stanford, CA 94305.
T. Hunkapiller, Division of Biology, California Institute of Technology, Pasadena, CA 91125.

loop. In addition to the similarity in amino acid sequence, computer predictions of potential β-sheet structure and hydrophobicity plots were extremely similar for the NH_2-terminal domain of L3T4 and for Ig light and heavy chain V regions. Twelve residues beyond the second cysteine of the V-like domain is a sequence (amino acids 102 to 110) that bears some similarity to Ig λ J (joining) segments (Fig. 2C). The placement of this sequence relative to the V-like domain suggests that it may be evolutionarily related to Ig J segments, although it is missing the central Gly-X-Gly, which is highly characteristic of Ig J sequences.

We found another region of the L3T4 protein between amino acids 140 and 175 (V′) that appears more weakly related to the COOH-terminal portion of Ig heavy and light chain V regions, including the second cysteine of the Ig disulfide loop (Fig. 2D). This sequence is most closely related to the NH_2-terminal V-like sequence of L3T4, suggesting an internal duplication. Although the significance of this short region of homology is unclear, it is followed by a sequence (amino acids 176 to 185) that is even more similar to λ light chain J segments than the sequence following the NH_2-terminal V-like domain (Fig. 2E). It is therefore possible that this region of the protein also evolved from an ancestor of Ig V regions but that it suffered a major deletion of its NH_2-terminal portion during the course of evolution. Even though both L3T4 and Lyt-2 are homologous to κ variable regions, they show little homology to each other except for the conserved residues present in most members of the Ig gene superfamily. Both this sequence divergence and the fact that L3T4 has undergone an internal duplication resulting in two V-like homology units (in contrast to the single such unit in Lyt-2) suggest that L3T4 and Lyt-2 have had very different evolutionary histories after splitting off from κ. The remainder of the external protein (connecting peptide) bears no significant homology to other known proteins except to the human CD4.

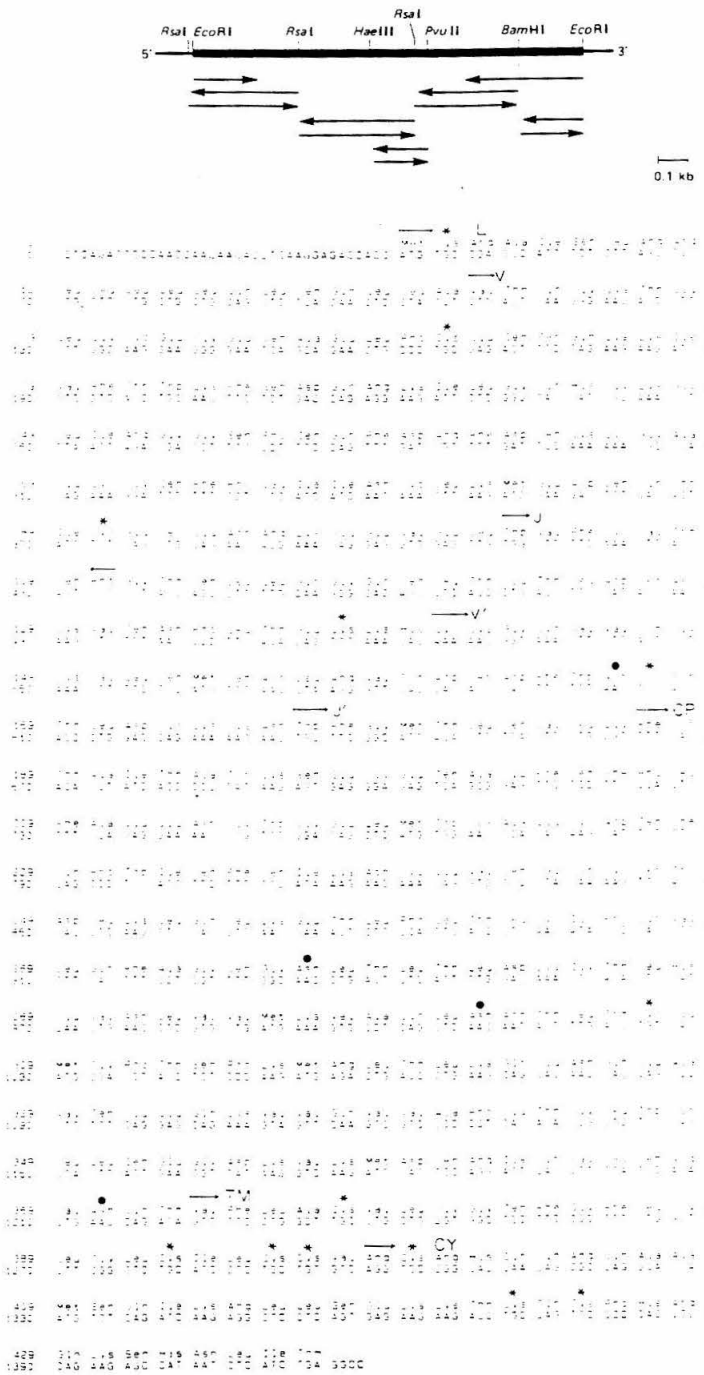


Fig. 1. Nucleotide and deduced amino acid sequence of L3T4 cDNA. The insert of cDNA clone pcL3T4-C7 was sequenced by the dideoxynucleotide chain termination method (*22*) with M13 vectors mp18 and mp19 (*23*) according to the strategy shown at the top. The closed box indicates the cDNA insert while the thinner lines indicate phage arms. The Eco RI sites at the 5′ and 3′ ends of the insert are the result of synthetic linker addition. The sequence of the 5′ untranslated region, leader, and first ten amino acids of the mature protein was determined similarly from M13 mp18 and mp19 subclones of a Sac I (5′) to Hinf I (3′) fragment of an overlapping cDNA clone, pcL3T4-14. This clone was found to be identical in sequence to pcL3T4-C7 in the region of overlap. The number of the first and last amino acid or nucleotide in each line is indicated in the left and right margins, respectively. Horizontal arrows indicate the start (or end) of the predicted protein domains. Cysteine residues are marked by asterisks and potential *N*-linked glycosylation sites by closed circles. Abbreviations: L, leader sequence (signal peptide); V and V′, sequences homologous to immunoglobulin variable regions; J and J′, sequences homologous to immunoglobulin joining segments; CP, connecting peptide; TM, transmembrane region; and CY, cytoplasmic tail.

The transmembrane region of L3T4 contains no charged residues and is only 54% homologous to the human CD4 transmembrane domain. In contrast, the transmembrane domain was the most highly conserved region (79%) between mouse Lyt-2 and human CD8 (*13*). Analysis of the L3T4 transmembrane region by a hydrophobic moment plot (*15*) predicts that this sequence penetrates the membrane as a single α-helix unassociated with any other transmembrane region. This result is consistent with biochemical data suggesting that L3T4 is a monomeric protein (*1*, *2*). The predicted cytoplasmic domain of L3T4 is highly charged, with an excess of 12 basic residues. This is similar to previous observations for Lyt-2, CD8, and CD4 (*11*–*13*). There is no significant homology between the cytoplasmic domain of L3T4 and that of other sequenced proteins (except its human homolog).

The pattern of expression of L3T4 mRNA in a variety of mouse cell lines and tissues was examined by Northern blot analysis in order to correlate mRNA expression with the known pattern of expression of the protein (Fig. 3A). The insert of L3T4 cDNA clone pcL3T4-C7 hybridized to a single major mRNA species of 3.7 kb in thymus, spleen, and lymph node. This same hybridizing mRNA was also seen in two thymoma cell lines known to express the protein (VL3/1 and KKT2). In contrast, no hybridizing mRNA was detectable in several other thymoma cell lines that do not express surface L3T4 (1112, MBL2, and R1.1R/TLIII 7X.6), in two B-cell lines (Bal 17 and 225), in an undifferentiated teratocarcinoma cell line (F9), in an Ia$^+$ macrophage line (WEHI-3), or in liver. Although not obvious in Fig. 3A, a small amount of the 3.7-kb mRNA was detectable in kidney on the original autoradiograph. It is possible that this mRNA represents contamination of the kidney RNA with RNA from L3T4$^+$ blood cells, but we cannot exclude the possibility of a low level of expression by endogenous renal cells. The thymus lane (lane 1) and, to a lesser extent, the VL3/1 lane (lane 10) in Fig. 3A are greatly overexposed and suggest the presence of a number of larger and smaller mRNA species. Several of these are believed to be artifacts related to compressions in the regions of the 28*S* and 18*S* ribosomal RNA (rRNA), possibly combined with a low level of degradation. These bands are not visible in repeated blots with different RNA preparations (Fig. 3B). However, the two largest bands are present on other blots and are probably partially spliced precursors, since both nuclear and cytoplasmic RNA are present.

Of particular interest was the finding that normal mouse brain contained a small amount of a 2.7-kb L3T4 transcript that was not convincingly detectable in any of the other tissues or cell lines examined by using multiple blots and different RNA preparations. However, we cannot exclude the possibility that small amounts of this transcript are present in other tissues at levels below that found in brain. Because an mRNA species comigrating with the brain 2.7-kb transcript was not clearly detectable in thymus, spleen, or lymph node, it is highly unlikely that the expression of this mRNA in brain is a result of contaminating L3T4$^+$ blood cells. This 2.7-kb transcript is found in the poly(A)$^+$ fraction of brain RNA and not in the poly(A)$^-$ fraction (Fig. 3C). We do not know which cells in the brain express this mRNA and whether they are of neuronal, glial, or possibly macrophage origin.

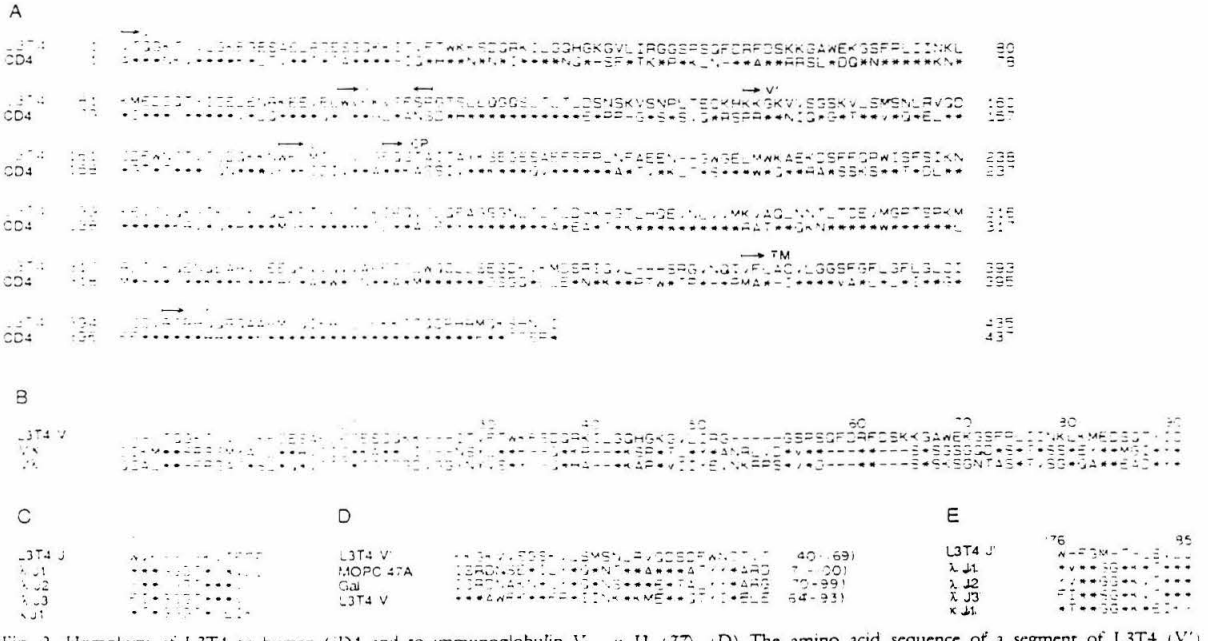


Fig. 2. Homology of L3T4 to human CD4 and to immunoglobulin V regions. (A) The amino acid sequence of L3T4 is compared with that of human CD4. Predicted protein domains of L3T4 are labeled as in Fig. 1. The numbers of the first and last amino acids in each line are indicated in the left and right margins, respectively. (B) The sequence of the NH_2-terminal domain of the mature L3T4 protein is compared with a mouse Ig κ (V-T1) V region (*24*) and a human Ig λ (Meg) V region (*25*). The numbers above refer to the amino acid positions in L3T4. (C) The amino acid sequence of a segment of L3T4 (residues 102 to 110) homologous to Ig J segments (L3T4 J) is compared with the sequences of mouse λ J1, J2, and J3 (*26*) and mouse κ J1 (*27*). (D) The amino acid sequence of a segment of L3T4 (V′) homologous to the COOH-terminal portion of Ig V regions is shown with the sequences of the homologous portion of two Ig heavy chain V regions, mouse MOPC 47A (*28*) and human GAL (*29*), and the homologous portion of the L3T4 NH_2-terminal V-like region (L3T4 V). The numbers in parentheses indicate the residues of each protein shown. (E) The sequence of a second J-like sequence of L3T4, J′ (residues 176 to 185) is compared with the sequences of three mouse λ J segments (*26*) and one mouse κ J segment (*27*). In all of these comparisons, asterisks indicate identical residues and dashes indicate gaps in the alignment.

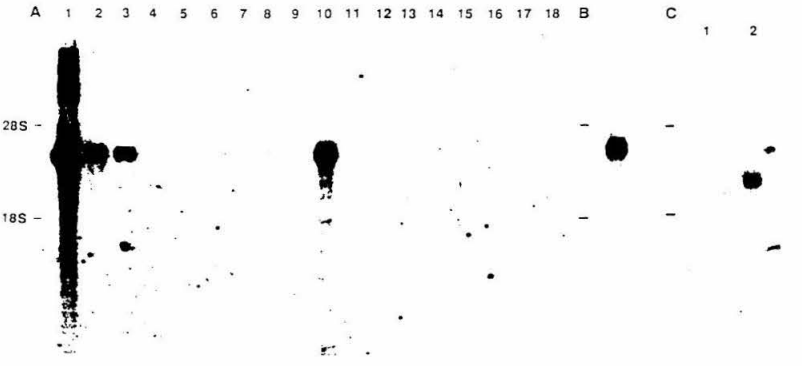


Fig. 3. Northern blot analysis of L3T4 mRNA expression. Total RNA was isolated from tissues and cell lines by the procedure of Chirgwin *et al.* (*30*). (A) Each RNA sample (10 μg) was subjected to electrophoresis through a 1.5% agarose gel containing 2.2*M* formaldehyde. The RNA was transferred to nitrocellulose (*31*) and the blot was hybridized as described (*13*). The probe consisted of a 1.3-kb insert of cDNA clone pcL3T4-C7 labeled with ^{32}P by random hexamer priming (*32*). The blot was washed as described (*13*) and exposed for 9 days at −70°C with an intensifying screen. RNA's were from the following sources: 1, thymus; 2, lymph node; 3, spleen; 4, brain; 5, kidney; 6, liver; 7, rat glial cell line C6 (A. Perlman, Stanford); 8, thymoma cell line KKT2 (I. Weissman, Stanford); 9, no sample; 10, thymoma cell line VL3/1 (I. Weissman, Stanford); 11, thymoma cell line 1112 (J. Allison, University of California, Berkeley); 12, thymoma cell line MBL2 (I. Weissman, Stanford); 13, thymoma cell line R1.1R/TLIII 7X.6 (R. Hyman, Salk Institute); 14, B-cell line Bal 17 (I. Weissman, Stanford); 15, B-cell line 225 (J. Allison, University of California, Berkeley); 16, F9 teratocarcinoma cell line; 17, macrophage cell line WEHI-3 (P. Jones, Stanford); and 18, WEHI-3 after 48 hours of induction with a γ-interferon–containing cell-free supernatant from concanavalin A–treated (48 hours) mouse spleen cells (*33*). All tissues and cell lines were of mouse origin except for the rat glial cell line C6. (B) Total thymus RNA (15 μg) was subjected to electrophoresis, blotted, and hybridized as in (A) except that the RNA was prepared independently. Exposure was for 5 hours at −70°C with an intensifying screen. (C) Total brain RNA isolated as in (A), but independently, was separated into poly(A)$^-$ (lane 1) and poly(A)$^+$ (lane 2) fractions by passage through and elution from an oligo(dT)-cellulose column (*34*) two times. Each fraction (9 μg) was subjected to electrophoresis, blotted, and hybridized as in (A). Exposure was for 18 hours at −70°C with an intensifying screen. The migration positions of 28*S* and 18*S* rRNA are marked in the left margin of each autoradiograph and were determined by ethidium bromide staining.

Since we have previously found that the mouse L3T4 cDNA cross-hybridizes with a single rat gene on genomic Southern blots (*16*), we examined a rat glial cell line, C6, for expression of RNA hybridizing to the mouse L3T4 cDNA. No cross-hybridizing mRNA was detectable (Fig. 3A). However results with a single cell line do not rule out the possibility that glial cells are the source of the mouse 2.7-kb transcript.

The L3T4 mRNA contains a very long 3′ untranslated region (estimated at 1.6 to 1.7 kb), so it was possible that the difference in mRNA lengths resulted from two alternative polyadenylation sites. If so, the encoded protein would be the same in both cases. However, S1 nuclease mapping indicates that the two mRNA's differ in protein coding sequence at the 5′ end (*17*). Since we have only detected a single gene for mouse L3T4, it seems most likely that the difference in the two mRNA's is the result of alternative mRNA splicing patterns. The typical 3.7-kb L3T4 transcript is also detectable in brain, but at a lower level. We do not know whether the same cells in brain express both mRNA species. It is possible that the small amount of 3.7-kb transcript in brain results from contamination with blood cells.

The homology of L3T4 to Ig V regions and the expression of an alternative mRNA form in brain led us to examine whether this gene rearranges in cells that express it. We therefore examined genomic DNA from mouse liver (nonexpressing), VL3/1 (a thymoma cell line expressing the 3.7-kb mRNA), and mouse brain (3.7-kb and 2.7-kb mRNA's) by Southern blot (*18*) analysis

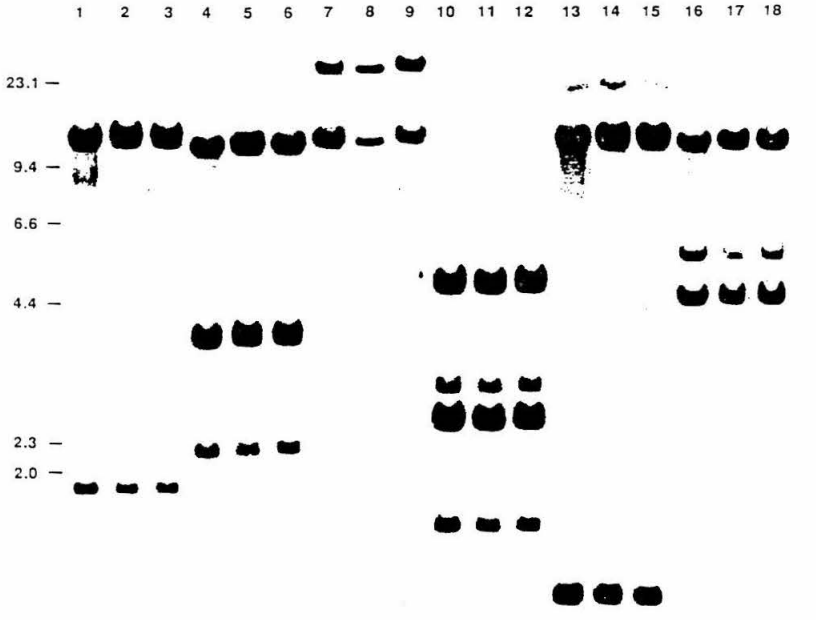


Fig. 4. Southern blot analysis of L3T4 gene in thymoma cell line, liver, and brain. Genomic DNA (8 μg) was digested with the indicated restriction endonuclease and subjected to electrophoresis through a 0.8% agarose gel. The DNA was transferred to nitrocellulose by the method of Southern (*18*) and the blot was hybridized, as previously described (*13*), to the insert of L3T4 cDNA clone pcL3T4-C7 labeled as described in the legend to Fig. 3. The blot was washed as described (*13*) and exposed overnight at −70°C with an intensifying screen. Enzymes used were (lanes 1 to 3) Bam HI; (lanes 4 to 6) Bgl II; (lanes 7 to 9) Hind III; (lanes 10 to 12) Pvu II; (lanes 13 to 15) Eco RI; and (lanes 16 to 18) Xba I. Sources of DNA were L3T4$^+$ mouse thymoma cell line VL3/1 (lanes 1, 4, 7, 10, 13, and 16), mouse liver (lanes 2, 5, 8, 11, 14, and 17), and mouse brain (lanes 3, 6, 9, 12, 15, and 18). Size markers are indicated in the left margin and represent Hind III fragments of phage λ run in parallel and stained with ethidium bromide.

after digestion with six different restriction enzymes (Fig. 4). We could detect no rearrangement in the L3T4 gene with the complete insert of pcL3T4-C7 as a probe. On the basis of the data from VL3/1 we conclude that no rearrangement is required for expression in T-lineage cells. Although we found no evidence for rearrangement in total brain DNA, we cannot rule out the possibility that a rearrangement occurs in fewer than 5% of brain cells, and that these are the cells that produce the smaller mRNA.

The expression of L3T4 mRNA in mouse brain is intriguing given the high frequency of central nervous system involvement in human acquired immune deficiency syndrome (AIDS) (*19*, *20*); the presence of human T-cell lymphotropic virus type III (HTLV-III), the retrovirus responsible for AIDS, in the brain of affected individuals (*19*); and the demonstrated role of human CD4 as a cellular receptor for HTLV-III (*21*). These observations suggest that CD4 is probably expressed in human brain at the protein level. The large evolutionary divergence that we have found between the external domains of L3T4 and CD4 may explain why HTLV-III does not infect mouse T cells. In any event, the tissue-specific expression of an alternative form of L3T4 mRNA suggests a yet undefined role for this molecule in brain.

REFERENCES AND NOTES

1. D. P. Dialynas *et al.*, *J. Immunol.* **131**, 2445 (1983).
2. D. P. Dialynas *et al.*, *Immunol. Rev.* **74**, 29 (1983).
3. S. L. Swain, *ibid.*, p. 129.
4. M. Okada and C. S. Henney, *J. Immunol.* **125**, 300 (1980); S. L. Swain, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 7101 (1981).
5. S. L. Swain, D. P. Dialynas, F. W. Fitch, M. English, *J. Immunol.* **132**, 1118 (1984); E. Nakayama, H. Shiku, E. Stockert, H. F. Oettgen, L. J. Old, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 1977 (1979); N. Hollander, E. Pillemer, I. L. Weissman, *J. Exp. Med.* **152**, 674 (1980).
6. H. R. MacDonald, A. L. Glasebrook, C. Bron, A. Kelso, J. C. Cerottini, *Immunol. Rev.* **68**, 89 (1982); I. L. Greenstein, J. Kappler, P. Marrack, S. J. Burakoff, *J. Exp. Med.* **159**, 1213 (1984).
7. T. R. Malek, C. Chan, L. H. Glimcher, R. N. Germain, E. M. Shevach, *J. Immunol.* **135**, 1826 (1985); P. Wassmer, C. Chan, L. Logdberg, E. M. Shevach, *ibid.*, p. 7688; I. Bank and L. Chess, *J. Exp. Med.* **162**, 1294 (1985).
8. G. S. Wood, N. L. Warner, R. A. Warnke, *J. Immunol.* **131**, 212 (1983); R. A. Moscicki, E. P. Amento, S. M. Krane, J. T. Kurnick, R. B. Colvin, *ibid.*, p. 743; W. A. Jeffries, J. R. Green, A. F. Williams, *J. Exp. Med.* **162**, 117 (1985).
9. P. Kavathas, V. P. Sukhatme, L. A. Herzenberg, J. R. Parnes, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 7688 (1984).
10. D. R. Littman, Y. Thomas, P. J. Maddon, L. Chess, R. Axel, *Cell* **40**, 237 (1985); V. P. Sukhatme, K. C. Sizer, A. C. Vollmer, F. Hunkapiller, J. R. Parnes, *ibid.*, p. 591.
11. P. J. Maddon *et al.*, *ibid.* **42**, 93 (1985).
12. R. Zamoyska, A. C. Vollmer, K. C. Sizer, C. W. Liaw, J. R. Parnes, *ibid.* **43**, 153 (1985).
13. H. Nakauchi, G. P. Nolan, C. Hsu, H. S. Huang, P. Kavathas, L. A. Herzenberg, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 5126 (1985).
14. D. Eisenberg, E. Schwarz, M. Komaromy, R. Wall, *J. Mol. Biol.* **179**, 125 (1984).
15. E. H. Field, B. Tourvieille, J. R. Parnes, unpublished results.
16. S. D. Gorman, B. Tourvieille, J. R. Parnes, unpublished results.
17. E. M. Southern, *J. Mol. Biol.* **98**, 503 (1975).
18. G. M. Shaw *et al.*, *Science* **227**, 177 (1985); J. A. Levy *et al.*, *Lancet* **1985-II**, 586 (1985); D. D. Ho *et al.*, *N. Engl. J. Med.* **313**, 1493 (1985).
19. W. D. Snider *et al.*, *Ann. Neurol.* **14**, 403 (1983); A. L. Belman *et al.*, *ibid.* **18**, 560 (1985); C. A. Carne *et al.*, *Lancet* **1985-II**, 1206 (1985).
20. A. G. Dalgleish *et al.*, *Nature (London)* **312**, 763 (1984); D. Klatzmann *et al.*, *ibid.*, p. 767; J. S. McDougal *et al.*, *J. Immunol.* **135**, 3151 (1985); J. A. Levy *et al.*, *Ann. Intern. Med.* **103**, 694 (1985); J. S. McDougal *et al.*, *Science* **231**, 382 (1986).
21. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977).
22. J. Messing, *Methods Enzymol.* **101**, 20 (1983).
23. W. Altenburger, M. Steinmetz, H. G. Zachau, *Nature (London)* **287**, 603 (1980); S. Bodary and B. Mach, *EMBO J.* **1**, 719 (1982).
24. J. W. Fett and H. F. Deutsch, *Biochemistry* **13**, 4102 (1974).
25. B. Blomberg, A. Traunecker, H. Eisen, S. Tonegawa, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 3765 (1981); B. Blomberg and S. Tonegawa, *ibid.* **79**, 530 (1982).
26. E. E. Max, J. G. Seidman, P. Leder, *ibid.* **76**, 3450 (1979).
27. E. A. Robinson and E. Appella, *J. Biol. Chem.* **254**, 11418 (1979).
28. S. Watanabe, H. U. Barnikol, J. Horn, J. Bertram, N. Hilschmann, *Z. Physiol. Chem.* **354**, 1505 (1973).
29. J. M. Chirgwin, A. E. Przybyla, R. J. MacDonald, W. J. Rutter, *Biochemistry* **18**, 5294 (1979).
30. P. S. Thomas, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 5201 (1980).
31. A. P. Feinberg and B. Vogelstein, *Anal. Biochem.* **132**, 6 (1983).

32. D. Paulnock-King, K. Sizer, Y. R. Freund, P. P. Jones, J. R. Parnes, *J. Immunol.* **135**, 632 (1985).
33. H. Aviv and P. Leder, *Proc. Natl. Acad. Sci. U.S.A.* **69**, 1408 (1972).
34. We thank D. O'Neill for editorial assistance. Supported by NIH grants AI11313 and GM34991; by the Ministère des Affaires Etrangeres-France, the Ligue Nationale Française contre le Cancer, and the Philippe Foundation (B.T.); by NCI postdoctoral fellowship 1 F32 CA07877-01 (S.D.G.); by a Bank of America–Giannini Foundation postdoctoral fellowship (E.H.F.); and by a John A. and George L. Hartford faculty fellowship award (J.R.P.).

6 June 1986; accepted 25 August 1986

VI.3 Major Histocompatibility Complex

Cell, Vol. 24, 125–134, April 1981, Copyright © 1981 by MIT

Three cDNA Clones Encoding Mouse Transplantation Antigens: Homology to Immunoglobulin Genes

Michael Steinmetz,* John G. Frelinger,* Douglas Fisher,* Tim Hunkapiller,* Dennis Pereira,† Sherman M. Weissman,† Hiroshi Uehara,‡ Stanley Nathenson,‡ and Leroy Hood*

* Division of Biology
California Institute of Technology
Pasadena, California 91125
† Department of Human Genetics
Yale University School of Medicine
New Haven, Connecticut 06510
‡ Departments of Microbiology and Immunology and Cell Biology
Albert Einstein College of Medicine
Bronx, New York 10461

Summary

We constructed cDNA libraries from poly(A)$^+$ RNA isolated from cell lines of two different inbred strains of mice, and screened the libraries with a cDNA clone encoding a human transplantation antigen. Three cDNA clones were identified, sequenced and found to encode amino acid sequences highly homologous to portions of a known mouse transplantation antigen. Comparison of the cDNA sequences of mouse transplantation antigens with the constant region domains of the mouse immunoglobulin μ gene reveals a striking homology, which suggests that the two genes share a common ancestor. Antibody genes undergo DNA rearrangements during B cell differentiation that are correlated with their expression. In contrast, DNA blots with these cDNA probes suggest that the genes for the transplantation antigens are not rearranged in the genomes of liver or embryo cells, which express these antigens, as compared with sperm cells, which do not express these antigens. In Bam HI–digested liver DNAs from different inbred strains of mice, 10–15 bands of hybridization were found. Accordingly, the genes encoding the transplantation antigens appear to constitute a multigene family with similar gene numbers in different mice.

Introduction

The major histocompatibility complex of mammals is a tightly linked cluster of genes encoding a variety of proteins involved in the immune response. One family of these molecules is denoted the transplantation or histocompatibility antigens because differences in these proteins can cause rapid graft rejection (Gorer, 1938; Gorer et al., 1948). Transplantation antigens are found on the cell surfaces of all mammalian somatic cells, albeit at varying concentrations. They play a fundamental role in T cell surveillance mechanisms, which recognize virally infected or in some cases neoplastically transformed cells. Indeed, cytotoxic T cells are restricted to killing cells that express both a foreign determinant, such as a viral antigen, and a self-transplantation antigen (for review see Shearer and Schmitt-Verhulst, 1977; Zinkernagel and Doherty, 1980).

The mouse is an ideal subject for the study of genes encoding transplantation antigens because detailed serological and genetic analysis of the mouse major histocompatibility (H-2) complex has been facilitated by the existence of inbred, congenic and recombinant strains (Klein, 1975; Snell et al., 1976). Some mice appear to have at least four major transplantation antigens, K, D, L and R—all encoded on chromosome 17 (Hansen et al., 1981). The K and D genes are approximately 0.5 centimorgans apart while L and R are closely linked to D, but their relative positions have not yet been determined. Individual alleles of the H-2 complex are linked together in a large number of distinct combinations called haplotypes. The haplotype of an inbred mouse strain is denoted by a small letter, for example, d for the inbred BALB/c strain. The transplantation antigens of BALB/c mice are denoted as the K^d, D^d, L^d and R^d molecules. Transplantation antigens are extremely polymorphic by serological analyses; for example, in the mouse there are at least 56 alleles at the K locus and 45 alleles at the D locus (Klein, 1979).

The transplantation antigens of mice and other species exhibit homologous structures comprising two polypeptide chains—an integral membrane glycoprotein of approximately 45,000 daltons that is noncovalently associated with a 12,000 dalton component, β_2–microglobulin (Vitetta and Capra, 1978). Hereafter we shall use the term transplantation antigen to refer only to the 45,000 dalton component. The gene for the 45,000 dalton component is encoded in the H-2 complex, whereas the gene for the mouse β_2–microglobulin has not yet been mapped.

The complete amino acid sequence of 346 residues for the K^b transplantation antigen has been determined (Martinko et al., 1980; Uehara et al., 1981). Partial amino acid sequences of other mouse K, D and L transplantation antigens demonstrate that they are approximately 80% homologous to one another (Nathenson et al., 1981). The extensive serological polymorphism among molecules encoded by different alleles is reflected in extensive amino acid substitutions (Maizels et al., 1978; Coligan et al., 1980; Nairn et al., 1980; Nathenson et al., 1981). In addition, amino acid sequence comparisons suggest that portions of the transplantation antigens may be homologous to immunoglobulin molecules (Orr et al., 1979; Strominger et al., 1980).

Two groups of investigators have cloned cDNA probes for human transplantation antigens (Ploegh et al., 1980; Sood et al., 1981). We report the characterization of three cDNA clones encoding three distinct

Cell
126

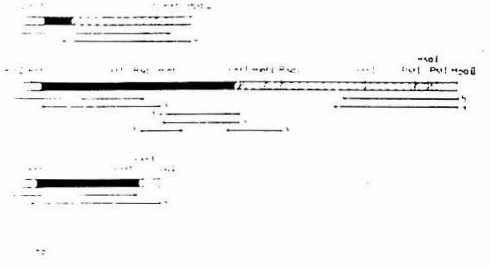


Figure 1. Partial Restriction Maps for the cDNA Clones pH-2I, pH-2II and pH-2III

Restriction sites were mapped on the cDNA inserts of the three plasmid DNAs by single and double restriction enzyme digestions in order to facilitate DNA sequence analysis. Arrows: extent and direction of DNA sequencing, 5' or 3' labeled ends. Open boxes: pBR322 sequences and G-C tails; filled boxes: coding sequences; hatched boxes: 3' untranslated regions

mouse transplantation antigens and demonstrate a significant homology relationship between a portion of the genes encoding the transplantation antigens and immunoglobulin constant region domains.

Results and Discussion

Three Mouse cDNA Clones Encode Molecules Which Are Highly Homologous to the K^b Transplantation Antigen

Using a human HLA cDNA clone (Sood et al., 1981) as a probe, we screened two cDNA libraries that were constructed from poly(A)$^+$ RNA from two mouse lymphoma cell lines. The C14 cell line, induced in a BALB/c (d haplotype) mouse by Abelson virus, overproduces the D^d molecule (Nairn et al., 1980), and the lymphoma cell line RDM-4 (k haplotype) overproduces the K^k antigen (Herrmann and Mescher, 1979). Two

Figure 2. DNA Sequences for the Inserts of the cDNA Clones pH-2I, pH-2II and pH-2III

The complete DNA sequences of the noncoding strands for the inserts of clones pH-2I and pH-2III are given together with the predicted amino acid sequences encoded by the cDNA clones. For clone pH-2II the sequence is given except for a stretch of about 170 bp in the 3' untranslated region. When compared with the completely sequenced K^b transplantation antigen (see Figure 3), clone pH-2I encodes amino acids 313 to 339, clone pH-2II codes for amino acids 167 to 352, and clone pH-2III encodes amino acids 63 to 160.

cDNA clones (pH-2I and pH-2II) were isolated from the C14 library and one cDNA clone (pH-2III) was isolated from the RDM-4 library. The approximate sizes of the inserts ranged between 300 and 1150 nucleotides: pH-2I, ~400 bp; pH-2II, ~1150 bp; and pH-2III, ~300 bp.

In order to characterize these clones, we determined the nucleotide sequences of the inserts for the three cDNA clones. Figure 1 gives partial restriction maps of the three cDNA clones and shows the DNA sequencing strategy employed in each case. Figure 2 gives DNA sequences obtained for these clones. These three DNA sequences were translated into amino acid sequences in all possible reading frames and then were compared with the fully determined amino acid sequence of the K^b molecule (Figure 3). Unambiguous homologies between the K^b molecule and the amino acid sequences were found in one reading frame for each of the three cDNA clones. Amino acid sequences in the correct reading frame are given in Figure 2.

The three isolated cDNA clones encode different portions of the transplantation antigen (Figure 3). Clone pH-2III codes for amino acids 63–160 and is 89% homologous to the K^b sequence. Clone pH-2II codes for amino acids 167–352, is 80% homologous to the K^b molecule and contains about 600 nucleotides of the 3′ untranslated region. Clone pH-2I codes for 27 amino acids at the C terminus, is 89% homologous to the K^b molecule and extends 274 nucleotides into the 3′ untranslated region.

We have compared the translated sequences for these three cDNA clones against the available protein sequence data for transplantation antigens derived from mice of the k and d haplotypes (Nairn et al., 1980; Rothbard et al., 1980; J. E. Coligan, personal communication). These comparisons allow us to exclude certain possibilities (Table 1) and to conclude that each of these cDNA clones encodes a distinct transplantation antigen, but no unambiguous assignments can be made, in part because of the paucity of amino acid sequence data available for the transplantation antigens (for example, no sequence data are available on the R^d molecule). Accordingly, it will be important to obtain protein sequence data as well as DNA data for subsequent gene and protein correlations in this system.

The pH-2II cDNA Clone Contains Repetitive DNA Sequences

When we used the pH-2II cDNA clone to analyze mouse DNA cleaved with Eco RI by Southern blot hybridization, this probe hybridized to a large number of genomic DNA fragments (Figure 4A). This type of hybridization indicates that the corresponding cDNA clone contains one or more repetitive elements (Steinmetz et al., 1980). To localize the repetitive element(s) on the cDNA clone, total mouse DNA was used as a probe against pH-2II DNA cleaved with various restriction enzymes, under conditions where only repetitive sequences will hybridize (Figure 4B). The repetitive sequence was mapped to the 3′ end of the pH-2II cDNA sequence. Indeed, restriction fragments from mouse DNA hybridized with two different intensities.

Table 1 cDNA Clones for Transplantation Antigens

Cell Line	Haplotype	Clone	Results from Protein and cDNA Sequence Comparisons
C14	d	pH-2II	Not K^d or D^d
		pH-2I	Different from clone pH-2II
RDM-4	k	pH-2III	Not K^k

The predicted amino acid sequence of clone pH-2II corresponds in 19 out of 19 positions that can be compared (Nairn et al., 1980) to the L^d amino acid sequence. It is therefore possible that pH-2II encodes the L^d molecule.

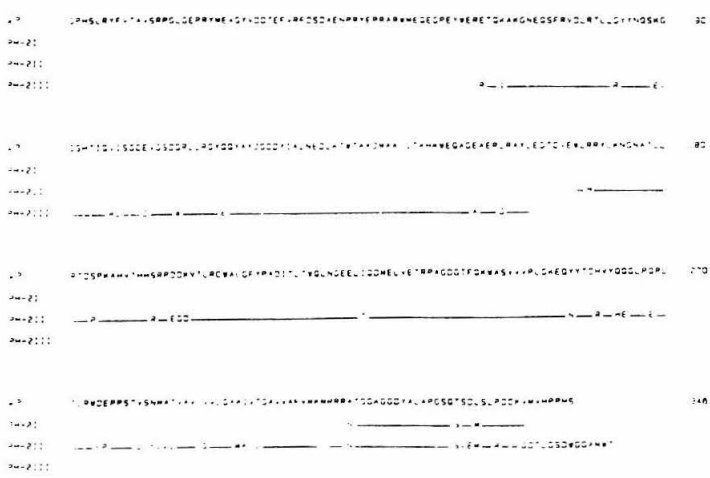


Figure 3. A Comparison of the Amino Acid Sequence of the Mouse K^b Molecule and the Translated Protein Sequences of the Three H-2 cDNA Clones

The mouse H-2K^b sequence has been published by Martinko et al. (1980) and Uehara et al. (1981). A gap of one amino acid at position 309 has been inserted into the predicted amino acid sequence of clone pH-2II to achieve maximum homology to the K^b molecule. The single-letter code has been used for amino acids: A, Ala; B, Asp or Asn; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; J, Glu or Gln; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr.

Solid line: identity to the K^b sequence.

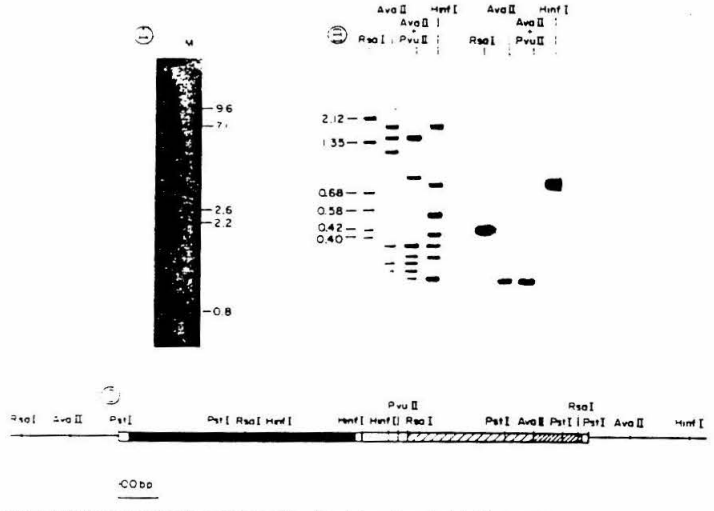


Figure 4 Localization of Repetitive DNA Sequences on the 3' Untranslated Region of Clone pH-2II

(A) The fragments from 10 μg of Eco RI-cleaved BALB/c mouse liver DNA (5×10^{-4} mole) were separated on a 0.6% agarose gel, transferred to a nitrocellulose filter and hybridized to clone pH-2II. The stringent wash was in 0.1 × SSC, 0.1% SDS at 68°C. Restriction fragments (7.5×10^{-16} mole per discrete fragment) of a pBR322 derivative were run in parallel (lane M) and served as molecular weight markers. Sizes are given in kb.

(B) The ethidium bromide–stained agarose gel (2%) on the left shows the restriction fragments obtained from the plasmid clone pH-2II upon digestion with the indicated nucleases. Fragment sizes are in kb. After blotting and hybridization with total mouse liver DNA, the autoradiogram on the right was obtained. Only sequences that are repetitive in mouse DNA are expected to hybridize under the conditions employed (Steinmetz et al., 1980).

(C) The map of the insert in clone pH-2II shows the localization of a moderately repetitive sequence between restriction sites for Rsa I and Ava II and a highly repetitive sequence between Ava II and the 3' end on the 3' untranslated region. Hatching: regions containing repetitive sequences; open boxes: G–C tails; filled boxes: coding region; thin line: pBR322 flanking sequences.

indicating the presence of a moderately repetitive sequence in the 3' untranslated region between the Rsa I and the Ava II site and a highly repetitive sequence between the Ava II site and the 3' end (Figure 4C). Sequence comparisons (not shown) indicate that these repetitive sequences do not fall into the Alu family described by Jelinek et al. (1980). To our knowledge this is the first example of repetitive sequence elements on the 3' untranslated region of a messenger RNA encoding a characterized protein.

The Genes Encoding Mouse Transplantation Antigens Do Show Convincing Homologies with Immunoglobulin Genes

Several features of transplantation antigens suggest that they may be evolutionarily related to immunoglobulins. First, the β_2-microglobulin polypeptide associated with transplantation antigens is ~30% homologous to immunoglobulin constant region domains, and accordingly has been denoted a "free immunoglobulin domain" (Peterson et al., 1972). Since immunoglobulin domains of light and heavy chains interact, perhaps β_2-microglobulin and the transplantation antigens interact through similar immunoglobulin-like domains. Second, immunoglobulins are comprised of multiple homology units or domains. These homology units, about 110 residues in length, are characterized by a centrally placed disulfide bridge spanning about 60 residues. The transplantation antigen also has two disulfide bridges, each spanning about 60 residues—that is, residues 101 to 164 and 203 to 259. Moreover, in human transplantation antigens, an 89 residue stretch of amino acid sequence including the second disulfide bridge (~181–270) shows a statistically significant homology (35%) to portions of immunoglobulin constant region domains at the protein level (Strominger et al., 1980). However, six sequence gaps must be inserted into the two sequences to obtain a 35% homology. One of the interesting questions that arises from these data is whether the protein homology arose by divergent (genes diverging from a common ancestral gene) or convergent (two independent genes converging toward a common protein sequence) evolution. This question cannot be answered at the protein level.

We used two computer programs to determine whether the genes for transplantation antigens were homologous to immunoglobulin genes. Initial analyses of the homology relationships of the H-2 cDNA clones to one another and to various immunoglobulin genes were done with the dot matrix computer program. Because this program cannot detect more distant and dispersed evolutionary relationships, we developed a second computer program, the best-fit matrix program (see Experimental Procedures for a description of these programs).

A best-fit matrix analysis of the pH-2II clone against the mouse immunoglobulin $C_\mu 4$ domain is given in Figure 5. The extended diagonal lines ("a" or "b") denote two homologous sequences located between nucleotides 119–364 of clone pH-2II and 1549–1809 in the $C_\mu 4$ domain of the μ gene. The homology is demonstrated at the DNA and protein levels in Figure 6. The two sequences are 51% (122 out of 237) homologous at the nucleotide level after the placement of two sequence gaps into each sequence. When

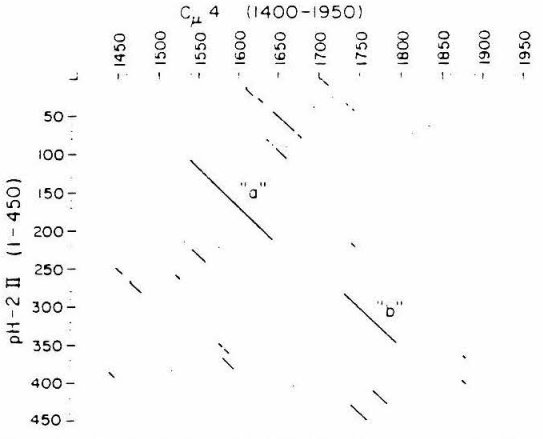


Figure 5. Homology between the pH-2II cDNA Sequence and the $C_\mu 4$ Domain of the Mouse Immunoglobulin μ Chain as Shown by the Best-fit Matrix Routine

We compared the pH-2II cDNA sequence from position 1 to 450 (vertical axis) with the DNA sequence of the $C_\mu 4$ domain from position 1400 to 1950 (horizontal axis) using the best-fit matrix program as described in Experimental Procedures. The numbering for the pH-2II sequence is the same as in Figure 2. The mouse μ chain sequence is from Kawakami et al. (1980) and nucleotide numbering for the μ gene is as published. Homologous regions in the analyzed sequences appear as a $-45°$ line in the matrix. The straight lines denoted by "a" and "b" indicate regions of significant homology between H-2 and the $C_\mu 4$ domain (see Figure 6). The significance of the homology is reinforced by the fact that the same regions of the pH-2II cDNA sequence are homologous with the $C_\mu 1$, $C_\mu 2$ and $C_\mu 3$ exons (results not shown).

the best-fit matrix is used to compare the same stretch of pH-2II sequence against the $C_\mu 1$, $C_\mu 2$ and $C_\mu 3$ domains, similar, although somewhat less extensive, homologies are noted (data not shown). Indeed, the same is true for the $C_{\gamma 1}$ and $C_{\gamma 2b}$ domains (data not shown). Moreover, this sequence in the pH-2II clone appears, if anything, to be more closely related to the C_μ domains than the C_α domains are to one another. The extended diagonal line "a" in Figure 5 represents a 72 base sequence (119–190) from clone pH-2II that is strikingly homologous to corresponding regions in the immunoglobulin domains $C_\mu 1$, $C_\mu 2$ and $C_\mu 4$ and is somewhat less homologous to that of the $C_\mu 3$ domain—53%, 52%, 60% and 36%, respectively (Figure 7). Within this 72 base stretch, there are 21 base positions conserved between the $C_\mu 1$, $C_\mu 2$ and $C_\mu 4$ sequences. Twenty of these 21 are conserved in clone pH-2II as well. This homology has been determined without placing any sequence gaps—thus permitting us to ask how frequently homologies at these levels would be seen if every possible stretch of 72 nucleotides in the pH-2II clone were compared against all possible blocks of 72 nucleotides in the mouse C_μ gene. About 1.3×10^6 such comparisons were made; the mean homology for these comparisons is 25% with a standard deviation of 5.4%. Thus the homologies exhibited by the comparisons of the pH-2II clone against the $C_\mu 1$, $C_\mu 2$, $C_\mu 3$ and $C_\mu 4$ domains in Figure 7 fell 5.2, 5.0, 2.0 and 6.6 standard deviations from the mean and, accordingly, are all highly statistically significant.

It is important to point out that the third base positions in codons are highly conserved in the comparisons shown in Figures 6 and 7. Indeed, 45% of the third base positions in the comparison between clone pH-2II and the $C_\mu 4$ domain in Figure 6 are conserved. Moreover, six third base positions are absolutely conserved in the 72 nucleotide sequence for pH-2II and the $C_\mu 1$, $C_\mu 2$ and $C_\mu 4$ domains (Figure 7), even though in three of these cases the amino acid is not conserved. These observations strongly suggest that the homologies between transplantation antigens and immunoglobulin arose by divergent rather than convergent evolution, because convergent evolution drives different genes to produce similar protein sequences without any selective pressures for the conservation of the third base positions in codons.

The homologous stretch of sequence in the transplantation antigens is located on the extracellular part of the molecule proximal to the cell membrane. Perhaps the reason this sequence is conserved is that this portion of the transplantation antigen must fold into the classic "antibody fold" (Poljak et al., 1973) in order to interact effectively in a noncovalent manner with β_2–microglobulin, which is folded into a similar configuration—much as the light and heavy chain domains of immunoglobulins interact with one another through a molecular complementarity in their similar antibody folds.

DNA Blots on Mouse Liver DNA Suggest That Transplantation Genes Are a Multigene Family

The homology between immunoglobulins, which are encoded by several large multigene families, and transplantation antigens raises questions about the multiplicity of genes encoding transplantation antigens. Indeed, several investigators have considered the possibility that the mouse genome contains many more transplantation antigen genes than required for the three or four polypeptides that are expressed in individual mice (Bodmer, 1973; Silver and Hood, 1976). In addition, there appear to be several other loci on chromosome 17 that encode gene products possibly related to transplantation antigens because of their similar size and association with β_2–microglobulin (for examples, the T cell differentiation antigens, such as TL, Qa1 and Qa2). Protein sequence data, however, are lacking for the T cell differentiation antigens, so that the degree of their sequence relatedness to transplantation antigens is unknown.

To obtain a rough estimate of the number of DNA sequences in the mouse genome that crosshybridize

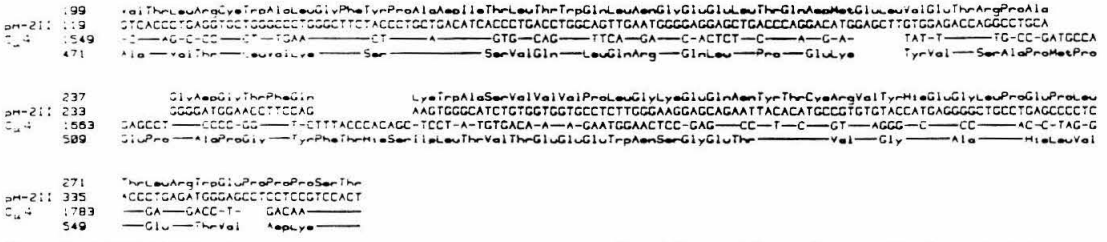


Figure 6. A Comparison between the DNA Sequences of Clone pH-2II and the Fourth Constant Region Domain of the Mouse Immunoglobulin μ Heavy Chain

A comparison is made between the pH-2II cDNA sequence from nucleotide positions 119 to 364 (Figure 2) and the fourth constant region exon ($C_\mu4$) of the μ gene from positions 1549 to 1809. This comparison includes the two homology regions identified in Figure 5 and shows that an overall homology of 51% can be achieved for the whole region with only two gaps in each sequence. The amino acid sequences encoded by the pH-2II and the $C_\mu4$ DNA sequences are given above and below the nucleotide sequences, respectively.
Solid line: $C_\mu4$ nucleotides and amino acids that are identical to the pH-2II sequence.

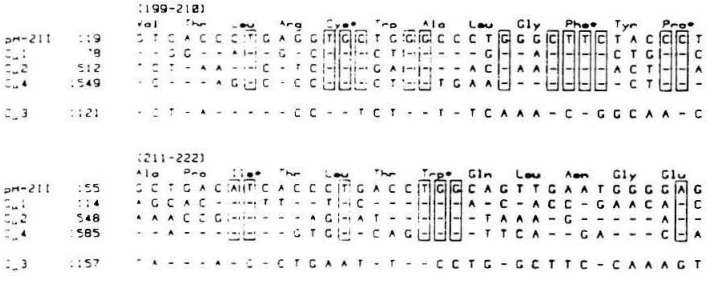


Figure 7. A Comparison between the DNA Sequences of the "a" Region of the pH-2II cDNA Clone and the Homologous Portions for the Immunoglobulin $C_\mu1$, $C_\mu2$, $C_\mu3$ and $C_\mu4$ Domains

The DNA sequence of clone pH-2II from nucleotide positions 119 to 190 (Figure 2) is compared with the DNA sequences of the four μ chain domains $C_\mu1$ to $C_\mu4$ (Kawakami et al., 1980) found to be homologous by the best-fit matrix comparison (see Figure 5). Dashes: identical nucleotides to the pH-2II sequence. Boxes: sequences that are conserved between pH-2II and the $C_\mu1$, $C_\mu2$ and $C_\mu4$ domains. The amino acid sequence encoded by pH-2II (positions 199 to 222) is given above the DNA sequence. Conserved amino acids are marked by an asterisk.

with our cDNA probes for transplantation antigens, we carried out Southern blot analyses of various mouse DNAs. Figure 8A shows a DNA blot analysis of a Bam HI digestion of BALB/c liver DNA with the pH-2II subclone. Figure 8B shows similar Bam HI analyses of sperm, embryo and liver DNA from BALB/c mice (d haplotype) and of liver DNAs from mice of the b, k and the recombinant a (K^kD^d) haplotypes, with the clone pH-2III as a probe. Several observations can be made. First, with both probes multiple bands of different intensities are detected. About 15 bands are identified with the pH-2II probe and about 12 with pH-2III probe. Second, the pH-2III and pH-2II probes, which represent the 5′ and 3′ areas of the coding sequence for transplantation antigens, hybridize to different, although probably overlapping, sets of restriction fragments. An analysis of five genomic clones crossreacting with these cDNA probes suggests that each putative transplantation antigen gene is cleaved a single time with the Bam HI enzyme and that in each case the pH-2II or pH-2III probes hybridize to just a single fragment (K. Moore, unpublished data).

The finding of multiple bands can be explained in several ways. First, the genes encoding transplantation antigens may be divided by many introns and thus many of the bands may represent a single gene. However, this explanation appears unlikely for many of the bands because of the detailed restriction analysis of the five distinct genomic clones mentioned above. Second, many of the bands may represent different genes, encoding transplantation antigens, as well as nonfunctional pseudogenes. Indeed, two factors may lead us to underestimate the number of distinct genes. Since the bands differ markedly in intensity, perhaps several of the heavy bands represent multiple distinct genes. In addition, our cDNA probes may not crossreact with all of the genes encoding transplantation antigens. It is unlikely, however, that there will be many more genes (bands), because variation of the stringency of hybridization does not alter the multiplicity of the bands. Accordingly, if all of these bands represent transplantation antigen genes, there could be 15 or more distinct genes. Third, perhaps the T cell differentiation antigens mentioned earlier are sufficiently homologous with transplantation antigens to crossreact with the cDNA probes employed. A detailed analysis of corresponding genomic clones and expressed proteins should allow us to differentiate between the latter two possibilities. We conclude that the genes encoding the transplantation antigens constitute a multigene family.

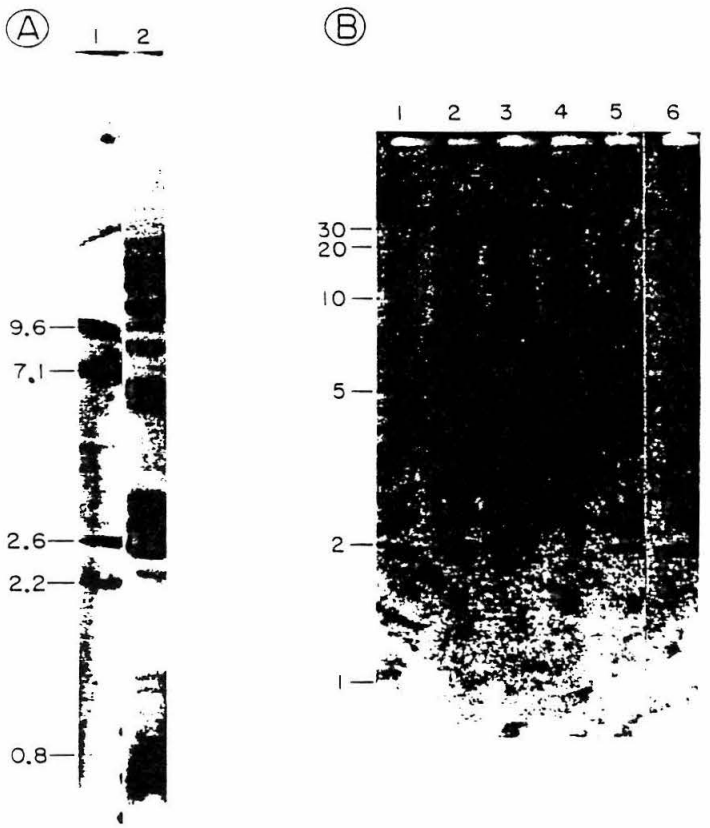


Figure 8. Southern Blot Hybridization of Mouse DNA with the pH-2II Subclone and the pH-2III cDNA Clone as Probes

(A) (Lane 2) 10 μg (5.6×10^{-18} mole) of BALB/c mouse liver DNA were cleaved with Bam HI to completion, separated on a 0.6% agarose gel and transferred to a nitrocellulose filter. Hybridization was with 50 ng/ml of a subclone containing the 688 bp Pst I–Pvn II fragment of pH-2II in pBR322 (see Figure 4). The final wash was in 0.1 × SSC, 0.1% SDS at 65°C. The hybridization markers in lane 1 are the same as in Figure 4, except that 5 × 10^{-18} mole per discrete fragment were used. (B) Each lane contained 3 μg of Bam HI–digested mouse DNA. Hybridization was carried out with 20 ng of the 287 bp Pst I fragment of clone pH-2III, which had been ligated to itself and labeled by nick translation. The final wash was in 1 × SSC, 0.1% SDS at 68°C. (Lanes 1, 2 and 3) sperm, embryo and liver DNA from BALB/c mouse (d haplotype); (lanes 4, 5 and 6) liver DNA from CBA (k haplotype), A/J (a haplotype) and B10 mice (b haplotype). λ DNA fragments were run in parallel and served as molecular weight markers. Sizes are in kb.

One final observation is that a single restriction enzyme polymorphism is seen between the haplotypes tested. This polymorphism occurs in the k haplotype, where a strongly hybridizing 5.4 kb band is seen instead of the 6.2 kb band found in the d haplotype (Figure 8B). Moreover, in the DNA from the recombinant a haplotype (K^kD^a) one finds the 6.2 kb band. This indicates that the polymorphism is encoded by the D region of the H-2 complex. In view of the extensive amino acid differences noted among the homologous transplantation antigens from different haplotypes (~5–10%), it is surprising to find only one restriction enzyme polymorphism. The simple interpretation of these observations is that the noncoding sequences of transplantation antigens are far more highly conserved than their coding regions.

DNA Blots on Sperm DNA Suggest That There Is No Rearrangement of the Genes Encoding Transplantation Antigens during the Differentiation of Murine Somatic Cells

The homology between the genes encoding immunoglobulins and transplantation antigens raises the interesting possibility that DNA rearrangements may be correlated with the expression of transplantation antigens, as they are associated with the expression of immunoglobulins (Brack et al., 1978; Seidman et al., 1979; Early et al., 1980). Transplantation antigens are not endogenously expressed by sperm cells (undifferentiated DNA), whereas they are expressed by virtually all somatic cells (differentiated DNA) (Klein, 1975). It is important to remember that at least four transplantation antigens are expressed in the BALB/c mouse, and if rearrangements were required for expression, all four genes should rearrange. By Southern blot hybridization, we analyzed Eco RI–, Bam HI– and Hind III–digested DNA from BALB/c sperm, embryo and liver DNA with the pH-2III clone and the single-copy subclone of pH-2II. Figure 8B shows the results obtained for the Bam digestion of these DNAs probed with the clone pH-2III. In all cases, no difference between differentiated and undifferentiated DNA was seen (Figure 8B lanes 1–3). Our failure to find differences between sperm and differentiated DNA by this method suggests that DNA rearrangements are unlikely within the gene segments of

transplantation antigens during differentiation. However, this is not unequivocal proof against DNA rearrangements, since all three restriction enzymes may cut between the putative recombination site and the portions of the gene for which we have a probe—a possibility rendered less likely by the fact there are at least four genes for transplantation antigens expressed in BALB/c mice.

The availability of cDNA clones for transplantation antigens will now allow the detailed study of the H-2 complex of chromosome 17 in the mouse, and will help to answer questions concerning the organization, expression and evolution of these H-2 genes.

Experimental Procedures

Materials

Restriction nucleases, T4 DNA kinase, exonuclease III and E. coli DNA polymerase I large fragment were purchased from New England Biolabs. Calf liver tRNA was obtained from Boehringer. Reverse transcriptase and E. coli DNA polymerase I, used for double-stranded cDNA synthesis, were gifts from J. Beard and M. Bond, respectively. Mouse sperm DNA was obtained from I. Weissman. The E. coli strain MC1061, originally provided by M. Casadaban, was obtained from T. Sargent. The mouse lymphoma cell line RDM-4 was provided by M. F. Mescher.

Preparation of Poly(A)$^+$ RNA

RDM-4 (H-2^k) and C14 (H-2^b) tumor cells were grown in Dulbecco's Modified Eagle's medium supplemented with 10% heat-inactivated fetal calf serum, 10 mM nonessential amino acids, 0.1 mg/ml glutamine, 50 μg/ml penicillin and 50 μg/ml streptomycin to a density of 4×10^5 and 2×10^6 cells/ml, respectively. Usually 10^9 cells were harvested, pelleted for 10 min at 1000 × g at 0°C and washed once in 50 ml 150 mM NaCl, 40 mM phosphate, pH 7.2. RNA was then isolated by precipitation of polysomes (Efstratiadis and Kafatos, 1976), or by the guanidinium thiocyanate procedure (Chirgwin et al., 1979) as described by Fyrberg et al. (1980). Poly(A)$^+$ RNA was selected by a single passage over oligo(dT)-cellulose (Efstratiadis and Kafatos, 1976).

Cloning of Double-stranded cDNA

Double-stranded cDNA was synthesized following published procedures (Friedman and Rosbash, 1977; Buell et al., 1978; Chang et al., 1978; Wickens et al., 1978; Efstratiadis and Villa-Komaroff, 1979) with minor modifications. About 2 μg of double-stranded cDNA was obtained from 10 μg of poly(A)$^+$-selected mRNA (estimated to be contaminated with poly(A)$^-$ RNA to about 50%). Its single-stranded length, as determined by alkaline agarose gel electrophoresis, was on the average 600 nucleotides. No size selection was employed. pBR322 DNA was cleaved with Pst I and the linearized molecule was purified by electrophoresis on a 1% agarose gel. It was recovered by electroelution in a dialysis bag (Smith, 1980) and was purified by BD-cellulose chromatography. About eight dGMP residues were added per 3′ end (Roychoudhury and Wu, 1980). To the double-stranded cDNA about 10–20 dCMP residues were added per 3′ end.

Equimolar amounts of tailed cDNA and tailed vector DNA were combined in 0.1 M NaCl, 10 mM Tris, 0.25 mM EDTA, pH 7.5 at a final DNA concentration of 1 μg/ml, heated to 64°C for 8 min and annealed by incubation in a water bath at 42–43°C. After 2 hr the water bath was switched off and cooled to room temperature overnight. Five micrograms of tRNA were then added as a carrier and the DNA was precipitated with ethanol. The pellet was washed with 80% ethanol, dried, dissolved in 0.1 M NaCl, 10 mM morpholinopropane sulfonic acid, pH 7.0 (10 μl per 300 ng of DNA) and stored on ice.

MC1061 E. coli cells were transformed following the procedure described by Kushner (1978), except that 150 ng of DNA were used per discrete transformation and L broth was used instead of Z broth. After transformation the cells were pelleted, resuspended in 0.2 ml of L broth and spread on nitrocellulose filters on plates containing 15 μg/ml tetracycline at a density of about 1000 colonies per filter. For both the C14 and RDM-4 libraries, 6000–8000 tetracycline-resistant colonies were obtained from 50 ng of double-stranded cDNA. Replica plating, amplification, storage and lysis of colonies on nitrocellulose filters were essentially as described by Hanahan and Meselson (1980).

Colony hybridization with a purified Pst I fragment containing a cloned HLA cDNA (Sood et al., 1981) was carried out in 3 × SSC, 1× Denhardt's solution (Denhardt, 1966), 10% dextran sulfate, 0.1% SDS, poly(rC), poly(rA), poly(rG), for 1 hr at 65°C and then at 55°C overnight. Filters were washed at 55°C in 3 × SSC, 1× Denhardt's solution, and then in 1 × SSC.

DNA Sequence Analysis

Restriction fragments were labeled at their 5′ ends with polynucleotide kinase and γ-^{32}P-ATP or at their 3′ ends with E. coli DNA polymerase I large fragment and α-^{32}P-dNTPs. The 3′ ends of Rsa I fragments were labeled by successive treatment with exonuclease III and E. coli DNA polymerase I large fragment, similar to the procedure described by Smith and Calvo (1980). The DNA was strand-separated and single strands were sequenced by the chemical degradation procedure (Maxam and Gilbert, 1980). For the G+A reaction the procedure described by Gray et al. (1978) was used. Reaction products were electrophoresed on 40 cm 20% acrylamide sequencing gels and 80 cm 5% and 8% acrylamide sequencing gels (Smith and Calvo, 1980).

DNA Blot Hybridization

DNA blots were prepared and hybridized with nick-translated probes as described (Schnell et al., 1980), except that the hybridization solution was supplemented with poly(rA), poly(rG) and poly(rC) (10 μg/ml each).

Computer Homology Programs

Initial sequence analysis of the H-2 cDNA clones was done with the aid of a dot matrix computer program, which belongs to DNAMST, a database and analysis system under development in this laboratory. This routine, presumably similar to those described by other laboratories (Efstratiadis et al., 1980; Hieter et al., 1980), looks at all possible alignments between all possible DNA sequence fragments of a desired length that can be generated from two sequences. A positive homology is scored if the percent homology of a comparison equals or exceeds a preset minimum. Homology within these parameters is displayed graphically as a dot. The coordinates of the dots are equivalent to the base positions of the lead bases of the two compared DNA segments on an X–Y field with axes represented by the two whole sequences. Regions of homology appear as various discrete patterns, but primarily as lines parallel to the diagonal of the matrix.

The H-2 clones were compared in this manner (data not shown) both with themselves and with each other, as well as against the mouse genomic C_{α}, $C_{\gamma 1}$ and $C_{\gamma 2b}$ genes. The 3′ flanking sequence also was compared to representative Alu family sequences.

Dot matrix-like routines have a serious limitation when used to compare sequences that are distantly related, and that may have significant homology that is distributed in a diffuse manner. Often the dot matrix program will increase the background to the point that real homology is obscured, if the length of the unit sequence compared or the percentage of homology required for a positive score is lowered. To get around this limitation, we used another routine that is graphically similar to the dot matrix, but differs in how it scores and evaluates homology between DNA fragments. The best-fit matrix routine establishes a homology score for each possible single base alignment between any two sequences. This score depends on the positive or negative score of the two base positions in question, as well as on that between the corresponding bases of the 5′ and 3′ neighboring sequences. The contribution to a given score by any flanking se-

quence homology is inversely proportional to its distance in sequence length from the aligned base pair. Each base of one DNA sequence is compared in this manner with all the base positions of the other DNA sequence and a score for each comparison is generated. A dot is then plotted, as it is with the dot matrix routine, to represent homology, but only at those matrix coordinates that equal the minimum score generated for a given base.

The results of this type of plot resemble a dot matrix with diagonally parallel lines representing regions of relative homology. When the pH-2II clone was compared to the C_α gene sequence in this manner, two lines had high homology scores ("a" and "b" in Figure 5) and the remainder of the shorter lines had lower scores. A detailed discussion of these programs is now in preparation (T. Hunkapiller et al., unpublished data).

Acknowledgments

We thank Michael Douglas for computer graphics. M. S. is the recipient of a fellowship from the Deutsche Forschungsgemeinschaft and D. F. and T. H. are NIH trainees. J. G. F. is supported by an NIH postdoctoral fellowship. This work was supported by grants from the National Institutes of Health.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received December 24, 1980; revised January 23, 1981

References

Bodmer, W. F. (1973). A new genetic model for allelism at histocompatibility and other complex loci: polymorphism for control of gene expression. Transplant. Proc. *5*, 1471–1475.

Brack, C., Hirama, M., Lenhard-Schuller, R. and Tonegawa, S. (1978). A complete immunoglobulin gene is created by somatic recombination. Cell *15*, 1–14.

Buell, G. N., Wickens, M. P., Payvar, F. and Schimke, R. T. (1978). Synthesis of full length cDNAs from four partially purified oviduct mRNAs. J. Biol. Chem. *253*, 2471–2482.

Chang, A. C. Y., Nunberg, J. H., Kaufman, R. J., Erlich, H. A., Schimke, R. T. and Cohen, S. N. (1978). Phenotypic expression in E. coli of a DNA sequence coding for mouse dihydrofolate reductase. Nature *275*, 617–624.

Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. and Rutter, W. J. (1979). Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. Biochemistry *18*, 5294–5299.

Coligan, J. E., Kindt, T. J., Nairn, R., Nathenson, S. G., Sachs, D. H. and Hansen, T. H. (1980). Primary structural studies of an H-2L molecule confirm that it is a unique gene product with homology to H-2K and H-2D antigens. Proc. Nat. Acad. Sci. USA *77*, 1134–1138.

Denhardt, D. T. (1966). A membrane-filter technique for the detection of complementary DNA. Biochem. Biophys. Res. Commun. *23*, 641–646.

Early, P., Huang, H., Davis, M., Calame, K. and Hood, L. (1980). An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: V_H, D and J_H. Cell *19*, 981–992.

Efstratiadis, A. and Kafatos, F. C. (1976). The chorion of insects: techniques and perspectives. In Methods in Molecular Biology, *8*, J. A. Last, ed. (New York: Dekker), pp. 1–124.

Efstratiadis, A. and Villa-Komaroff, L. (1979). Cloning of double-stranded cDNA. In Genetic Engineering: Principles and Methods, *1*, J. K. Setlow and A. Hollaender, eds. (New York: Plenum), pp. 15–36.

Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C. and Proudfoot, N. J. (1980). The structure and evolution of the human β-globin gene family. Cell *21*, 653–668.

Friedman, E. Y. and Rosbash, M. (1977). The synthesis of high yields of full-length reverse transcripts of globin mRNA. Nucl. Acids Res. *4*, 3455–3471.

Fyrberg, E. A., Kindle, K. L., Davidson, N. and Sodja, A. (1980). The actin genes of Drosophila: a dispersed multigene family. Cell *19*, 365–378.

Gorer, P. A. (1938). The antigenic basis of tumour transplantation. J. Pathol. Bacteriol. *47*, 231–252.

Gorer, P. A., Lyman, S. and Snell, G. D. (1948). Studies on the genetic and antigenic basis of tumour transplantation: linkage between a histocompatibility gene and "fused" in mice. Proc. R. Soc. (Lond.) B *135*, 499–505.

Gray, C. P., Sommer, R., Polke, L., Beck, E. and Schaller, H. (1978). Structure of the origin of DNA replication of bacteriophage fd. Proc. Nat. Acad. Sci. USA *75*, 50–53.

Hanahan, D. and Meselson, M. (1980). Plasmid screening at high colony density. Gene *10*, 63–67.

Hansen, T. H., Ozato, K., Melino, M. R., Coligan, J. E., Kindt, T. J., Jandinski, J. J. and Sachs, D. H. (1981). A gene cluster in the *H-2D* region: evidence in two haplotypes for at least three *D* region-encoded molecules. D, L, and R. J. Immunol., in press.

Herrmann, S. H. and Mescher, M. F. (1979). Purification of the H-2K^k molecule of the murine major histocompatibility complex. J. Biol. Chem. *254*, 8713–8716.

Hieter, P. A., Max, E. E., Seidman, J. G., Maizel, J. V., Jr. and Leder, P. (1980). Cloned human and mouse kappa immunoglobulin constant and J region genes conserve homology in functional segments. Cell *22*, 197–107.

Jelinek, W. R., Toomey, T. P., Leinwand, L., Duncan, C. H., Biron, P. A., Choudary, P. V., Weissman, S. M., Rubin, C. M., Houck, C. M., Deininger, P. L. and Schmid, C. W. (1980). Ubiquitous, interspersed repeated sequences in mammalian genomes. Proc. Nat. Acad. Sci. USA *77*, 1398–1402.

Kawakami, T., Takahashi, N. and Honjo, T. (1980). Complete nucleotide sequence of mouse immunoglobulin μ gene and comparison with other immunoglobulin heavy chain genes. Nucl. Acids Res. *8*, 3933–3945.

Klein, J. (1975). Biology of the Mouse Histocompatibility-2 Complex (Berlin: Springer-Verlag).

Klein, J. (1979). The major histocompatibility complex of the mouse. Science *203*, 516–521.

Kushner, S. (1978). An improved method for transformation of Escherichia coli with Col E1 derived plasmids. In Genetic Engineering, H. W. Boyer and S. Nicosia, eds. (Amsterdam: Elsevier/North Holland Biomedical Press), pp. 17–23.

Maizels, R. M., Frelinger, J. A. and Hood, L. (1978). Partial amino acid sequences of mouse transplantation antigens. Immunogenetics *7*, 425–444.

Martinko, J. M., Uehara, H., Ewenstein, B. M., Kindt, T. J., Coligan, J. E. and Nathenson, S. G. (1980). Primary structure of murine major histocompatibility complex alloantigens: completion of the sequence of the amino-terminal 284 residues of H-2K^b. Biochemistry *19*, 6188–6193.

Maxam, A. M. and Gilbert, W. (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. Meth. Enzymol. *65*, 499–560.

Nairn, R., Nathenson, S. G. and Coligan, J. E. (1980). Isolation, characterization and amino acid sequence studies of the cyanogen bromide fragments of the H-2D^d glycoprotein. Eur. J. Immunol. *10*, 495–503.

Nathenson, S. G., Uehara, H., Ewenstein, B. M., Kindt, T. J. and Coligan, J. E. (1981). Primary structural analysis of transplantation antigens of the murine H-2 major histocompatibility complex. Ann. Rev. Biochem., in press.

Orr, H. T., Lancet, D., Robb, R. J., Lopez de Castro, J. A. and Strominger, J. L. (1979). The heavy chain of human histocompatibility antigen HLA-B7 contains an immunoglobulin-like region. Nature *232*, 266–270.

Peterson, P. A., Cunningham, B. A., Berggard, I. and Edelman, G. M. (1972). β_2-Microglobulin—a free immunoglobulin domain. Proc. Nat. Acad. Sci. USA *69*, 1697–1701.

Ploegh, H. L., Orr, H. T. and Strominger, J. L. (1980). Molecular cloning of a human histocompatibility antigen cDNA fragment. Proc. Nat. Acad. Sci. USA *77*, 6081–6085.

Poljak, R. J., Amzel, L. M., Avey, H. P., Chen, B. L., Phizackerley, R. P. and Saul, F. (1973). Three-dimensional structure of the Fab' fragment of a human immunoglobulin at 2.8 Å resolution. Proc. Nat. Acad. Sci. USA *70*, 3305–3310.

Rothbard, J. B., Hopp, T. P., Edelman, G. M. and Cunningham, B. A. (1980). Structure of the heavy chain of the H-2K^k histocompatibility antigen. Proc. Nat. Acad. Sci. USA *77*, 4239–4243.

Roychoudhury, R. and Wu, R. (1980). Terminal transferase-catalyzed addition of nucleotides to the 3' termini of DNA. Meth. Enzymol. *65*, 43–62.

Schnell, H., Steinmetz, M., Zachau, H. G. and Schechter, I. (1980). An unusual translocation of immunoglobulin gene segments in variants of the mouse myeloma MPC11. Nature *286*, 170–173.

Seidman, J. G., Max, E. E. and Leder, P. (1979). A K-immunoglobulin gene is formed by site specific recombination without further somatic mutation. Nature *280*, 370–375.

Shearer, G. M. and Schmitt-Verhulst, A. M. (1977). Major histocompatibility complex restricted cell-mediated immunity. Adv. Immunol. *25*, 55–91.

Silver, J. and Hood, L. (1976). Preliminary amino acid sequences of transplantation antigens: genetic and evolutionary implications. In Contemporary Topics in Molecular Immunology, *5*, H. N. Eisen and R. A. Reisfeld, eds. (New York: Plenum Publishing Corporation), pp. 35–68.

Smith, D. R. and Calvo, J. M. (1980). Nucleotide sequence of the E. coli gene coding for dihydrofolate reductase. Nucl. Acids Res. *8*, 2255–2274.

Smith, H. O. (1980). Recovery of DNA from gels. Meth. Enzymol. *65*, 371–380.

Snell, G. D., Dausset, J. and Nathenson, S. (1976). Histocompatibility. (New York: Academic Press).

Sood, A. K., Pereira, D. and Weissman, S. M. (1981). Isolation and partial nucleotide sequence of a cDNA clone for human histocompatibility antigen HLA-B by use of an oligodeoxynucleotide primer. Proc. Nat. Acad. Sci. USA, in press.

Steinmetz, M., Hochtl, J., Schnell, H., Gebhard, W. and Zachau, H. G. (1980). Cloning of V region fragments from mouse liver DNA and localization of repetitive DNA sequences in the vicinity of immunoglobulin gene segments. Nucl. Acids Res. *8*, 1721–1729.

Strominger, J. L., Orr, H. T., Parham, P., Ploegh, H. L., Mann, D. L., Bilofsky, H., Saroff, H. A., Wu, T. T. and Kabat, E. A. (1980). An evaluation of the significance of amino acid sequence homologies in human histocompatibility antigens (HLA-A and HLA-B) with immunoglobulins and other proteins, using relatively short sequences. Scand. J. Immunol. *11*, 573–592.

Uehara, H., Coligan, J. E. and Nathenson, S. G. (1981). Isolation and sequence analysis of membrane and hydrophilic segments of the H-2K^b alloantigen. Biochemistry, in press.

Vitetta, E. S. and Capra, J. D. (1978). The protein products of the murine 17th chromosome: genetics and structure. Adv. Immunol. *26*, 147–193.

Wickens, M. P., Buell, G. N. and Schimke, R. T. (1978). Synthesis of double-stranded DNA complementary to lysozyme, ovomucoid and ovalbumin mRNAs. J. Biol. Chem. *253*, 2483–2495.

Zinkernagel, R. M. and Doherty, P. C. (1980). MHC-restricted cytotoxic T cells: studies on the biological role of polymorphic major transplantation antigens determining T-cell restriction—specificity, function, and responsiveness. Adv. Immunol. *27*, 51–177.

Reprint Series
17 December 1982, Volume 218, pp. 1229–1232

SCIENCE

DNA Sequence of the Gene Encoding the E_α Ia Polypeptide of the BALB/c Mouse

Janet McNicholas, Michael Steinmetz, Tim Hunkapiller, Patricia Jones, and Leroy Hood

DNA Sequence of the Gene Encoding the E_α Ia Polypeptide of the BALB/c Mouse

Abstract. *A 3.4-kilobase DNA fragment containing the gene coding for the E_α chain of an Ia (I region–associated) antigen from the BALB/c mouse has been sequenced. It contains at least three exons, which correlate with the major structural domains of the E_α chain—the two external domains $\alpha 1$ and $\alpha 2$, and the transmembrane-cytoplasmic domain. The coding sequence of the mouse E_α gene shows striking homology to its human counterpart at the DNA and protein levels. The translated $\alpha 2$ exon demonstrates significant similarity to β_2-microglobulin, to immunoglobulin constant region domains, and to certain domains of transplantation antigens. These observations and those of others suggest that the Ia antigen, transplantation antigen, and immunoglobulin gene families share a common ancestor.*

The major histocompatibility complex (MHC) of the mouse encodes several families of cell-surface glycoproteins which regulate various aspects of immune responsiveness (*1*, *2*). Certain products of the class I genes, the transplantation antigens, serve as restricting elements in T cell immunosurveillance. The products of the class II genes, the I region–associated (Ia) antigens, play a fundamental role in determining the effectiveness of cell-cell interactions between regulatory T cells, B cells, and macrophages. A variety of recent data from genetic, functional, and biochemical studies suggest that the Ia antigens are the products of the Ir (immune response) genes, which control the ability of an animal to respond to synthetic and naturally occurring antigenic determinants (*3*–*7*).

Two Ia antigens, I-A and I-E, have

SCIENCE, VOL. 218, 17 DECEMBER 1982 0036-8075/82/1217-1229$01.00/0

been defined serologically and biochemically [reviewed in (*8*)]. Each is composed of an α and a β polypeptide chain, denoted A_α and A_β for I-A antigens and E_α and E_β for I-E antigens. The class II β chains appear to be quite polymorphic, whereas their α chain counterparts are less so. The set of alleles at the MHC loci of a particular inbred strain of mouse is denoted its haplotype. For example, the BALB/c mouse, whose genes we are studying, is of the d haplotype, and its class II genes are denoted A^d_α, A^d_β, E^d_α, and E^d_β.

Protease digestion studies of human α and β chains, protein sequence of a human β chain, as well as DNA sequence analyses of human class II complementary DNA (cDNA) clones, suggest that both the α and β polypeptides are divided into two external domains each of approximately 90 residues (α1 and α2 and β1 and β2), a transmembrane region of about 30 residues, and a cytoplasmic region of about 15 residues (*9–13*). Studies on the genomic organization of the class II genes are in progress in several laboratories. We now report the DNA sequence of the E_α gene of the BALB/c mouse.

We isolated the mouse E_α gene by screening a cosmid library constructed from BALB/c sperm DNA with a human DR_α (D-related) chain cDNA clone (*14*). A 3.4-kilobase (kb) Sal I fragment that hybridized to the human cDNA probe was isolated from cosmid 32.1 and cloned in both orientations into M13mp8; its nucleotide sequence was determined by the dideoxy chain termination method. A novel sequencing strategy adapted from the method of Frischauf *et al.* (*15*) was employed to generate an ordered series of subclones with increasing lengths of deletions starting at one end of the insert by deoxyribonuclease I digestion (*16*) (Fig. 1). The DNA sequence of the 3.4-kb fragment containing the E_α gene is given in Fig. 2.

The exons of the mouse E_α gene were identified on the basis of their homology to cDNA sequences of human DR_α clones (*13, 17*) as well as to the amino acid sequences of human DR_α chains and mouse E_α sequences (*18*). Three exons were identified. These exons encode the α1 domain (codons 3 to 84), the α2 domain (codons 85 to 178), and a domain which includes both the transmembrane and cytoplasmic regions (codons 179 to 230). A termination codon in phase with this reading frame is found at codon 231. Intervening sequences split the codons 3, 85, and 179 between base positions 1 and 2. Intervening sequences also split the first and second bases of codons in the genes encoding class I molecules, immunoglobulins, and β_2-microglobulin (*19–21*). The bases GT and AG are found at the 5′ and 3′ ends of each intron, respectively, as in virtually all other eukaryotic genes.

The three exons of the E_α gene contain the entire coding sequence except for the first two amino acids and a presumed leader peptide. Thus the E_α gene is split into at least four exons. Translation of the DNA sequence of 1480 bases 5′ to the α1 exon in all three reading frames did not reveal a hydrophobic stretch of amino acids beginning with a methionine and ending with isoleucine and lysine, which are the first two amino acids of the α1 domain of the E_α polypeptide from a mouse of the k haplotype (*18*). It appears

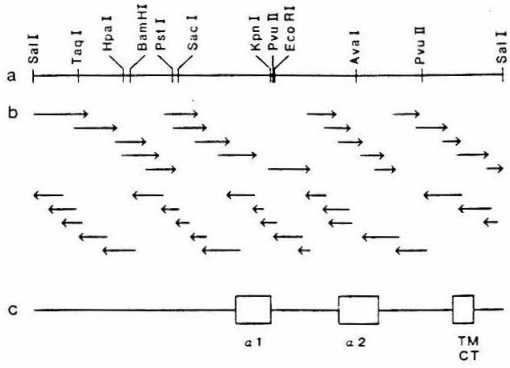


Fig. 1. The organization and sequencing strategy for the E_α gene. (a) A restriction map of the 3.4-kb Sal I fragment containing exons of the E_α gene which hybridized to the human DR_α probe (*14, 17*). The 3.4-kb fragment was subcloned into the Sal I site of pBR325 for mapping by single and double digests with the restriction enzymes indicated. This map was confirmed by computer analysis of restriction enzyme sites in the DNA sequence of this fragment shown in Fig. 2. (b) Sequence strategy for the E^d_α gene. Each arrow represents the sequence of an M13 clone. The 3.4-kb fragment from cosmid 32.1 (*14*) was cloned in both orientations into M13 mp8 to give two parental subclones, mp8-32.11 and mp8-32.12. Each parental subclone was then used to generate a series of overlapping subclones by a deoxyribonuclease deletion technique (*16*). (c) Organization of the E^d_α gene. Exons are represented by boxes and introns by lines. The 3.4-kb fragment contains one exon encoding most of the first protein domain, α1 (amino acids 3 to 84), a second exon encoding the second protein domain, α2 (amino acids 85 to 178), and a third exon that contains coding sequence for amino acids 179 to 230 and includes the transmembrane and cytoplasmic protein domains (*TM-CT*).

Table 1. Homology comparisons of the α2 exon of the E^d_α gene with class II, class I, immunoglobulin, β_2-microglobulin, and Thy-1 sequences. Sequences used to determine percent homology with the α2 exon of E^d_α were as follows: the α2 exon of the human DR_α gene (*13, 22*); the β_2 exon of the human pDR-β-1 cDNA clone (*12*); the β_2 domain of a human DR_β chain (*11*); the α3 exons of the mouse $H\text{-}2L^d$ gene (*28*), $H\text{-}2K^b$ gene (*29*), the Qa (27.1) pseudogene (*23*) and a human HLA gene (*30*); immunoglobulin constant regions (*31–34*); β_2-microglobulin (*21, 35*), and Thy-1 (*36*). Percent homology is based on alignments of the compared sequences given by best-fit matrix analysis (*23*). Alignments are based solely on amino acid sequence comparisons. Each percentage involves the full length of the α2 domain or exon of E_α (94 amino acids and 282 bases, respectively) and a comparable length sequence from the other protein or gene being compared. Insertions and deletions for alignment were minimized, and percentages given are only the minimum values. Percentages given are statistically significant. Each was compared to an expected percent homology and standard deviation for any random alignment of sequence fragments the length of α2 from E_α and the test sequence. The expected value is the mean homology of all possible fragments the appropriate length from E_α to all possible fragments of the same length from either the test sequence or some other unrelated sequence. Each comparison is four or more standard deviations from the expected homology.

Gene name	Percent homology		Insertions and deletions*
	Protein	DNA	
DR_α	81	82	0
pDR-β-1	32	46	0
DR_β chain†	32	N.A.‡	1
$H\text{-}2L^d$	24	37	0
$H\text{-}2K^b$	23	38	0
Qa (27.1)	23	37	0
HLA	26	38	0
$C_\alpha(C_H2)$	21	38	1
$C_\alpha(C_H4)$	22	44	1
$C_{\alpha 2}(C_H1)$	27	40	2
C_α	26	43	3
C_γ	30	40	2
β_2-microglobulin (mouse)	32	42	1
β^2-microglobulin (human)	31	43	1
Thy-1 chain†	20	N.A.‡	3

*Indicates the fewest number of insertions, deletions, or both, needed to align sequences for comparison. †These data are from protein sequences; no DNA sequences are available. ‡Sequence not available.

that the exon encoding the leader sequence and first two residues of E_α is separated from the α1 exon by an intervening sequence longer than ~1.5 kb.

About 250 base pairs (bp) of DNA has been sequenced 3′ to the third exon, and no polyadenylation signal has yet been observed. Indeed, we have compared the DNA sequence downstream of the termination codon of the E_α gene to the sequence of the 3′ untranslated region of the human DR_α chain cDNA clone (*13*, *17*) using a computer homology search program. Except for 13 bp immediately adjacent to the termination codon, no significant homology was found for about 250 bp of sequence that could be compared up to the end of the 3.4-kb Sal I fragment. The gene for DR_α, the human homolog of E_α, contains at least four exons for the α1 domain, α2 domain, the transmembrane together with the cytoplasmic domain, and the 3′ untranslated region. The presumptive exon encoding the leader peptide has not been identified and is located more than 0.4 kb upstream of the α1 exon (*22*). Thus, the structure of the mouse E_α gene is in complete agreement with the structure of the human DR_α gene.

The E_α gene appears, by sequence analysis, to be a functional gene in that it lacks any obvious elements that would render it a pseudogene, such as termination codons or inappropriate reading frameshifts. In addition, as mentioned, all of the exon-intron boundaries have the consensus upstream or downstream RNA-splicing signals.

Several lines of evidence indicate that the class II gene we have described is the E_α gene. First, the translated sequence of the E_α gene is identical to the NH_2-terminal 34 residues of the E_α polypeptide isolated from C3H mice except for one residue, which may reflect genetic polymorphism (*18*). Of course, the NH_2-terminal two residues are missing from our sequence. Second, the E_α coding region shows a 79 percent DNA sequence homology and 75 percent protein sequence homology to its human counterpart, DR_α (*13*, *14*, *22*). A comparison, exon by exon, indicates that the α1 exons of E_α and DR_α are as similar to each other as are the α2 exons to each other, 81 percent and 82 percent, respectively, at the DNA level and 79 percent and 81 percent, respectively, at the protein level. The transmembrane cytoplasmic exons show less homology (73 percent at the DNA level and 57 percent at the protein level). However, a comparison at the level of amino acid functional types (nonpolar, polar, basic, and acidic) gives a match of 88 percent, indicating strong

Fig. 2. Nucleotide sequence of the 3.4-kb Sal I fragment containing the E_α^d gene. The amino acid translation of each exon is given above the nucleotide sequence. Certain amino acid residues are encoded by two exons and are so indicated by splits in the three-letter amino acid code. Abbreviations: A, adenine; C, cytosine; G, guanine; T, thymine; Arg, arginine; Asn, asparagine; Asp, aspartic acid; Cys, cysteine; Gln, glutamine; Glu, glutamic acid; Gly, glycine; His, histidine; Ile, isoleucine; Leu, leucine; Lys, lysine; Met, methionine; Phe, phenylalanine; Pro, proline; Ser, serine; Thr, threonine; Trp, tryptophan; Tyr, tyrosine; and Val, valine.

conservation for function. The direct confirmation that this class II gene is indeed the E_α gene will await gene transfer and expression studies.

We have used a graphically displayed computer routine termed the best-fit matrix analysis (*23*) to analyze possible similarities between the DNA and protein sequences of the E_α gene exons and between these exons and other class II genes, class I genes, immunoglobulin genes, and Thy-1 antigen. Such analyses showed no significant similarities between the different domains of E_α or between the α1 and transmembrane-cytoplasmic domains of E_α and anything other than the same regions of the DR_α cDNA. However, similarity alignment was possible between an area of each of the tested sequences and the α2 domain of E_α. Table 1 lists the sequences compared and the percent homology of the aligned regions to the α2 domain of E_α at both the protein and DNA levels.

The α2 domain of E_α has significant similarity to "homology unit" (*9*) sequences of the genes listed, a sequence associated with the "antibody fold" tertiary structure of antibody domains (Table 1). This observation has been made by several other groups analyzing cDNA (*13*, *17*) or genomic (*22*) clones. Of the comparisons made, perhaps the most interesting is that to β_2-microglobulin. Not only is β_2-microglobulin as similar in sequence to any of the class II α2 and β2 domains as these are to each other, but it is strikingly similar in genomic organization to the E_α and DR_α genes. Like these two genes, β_2-microglobulin has its leader peptide and first two codons separated from the main protein coding sequence by a very large intervening sequence (2.8 kb). Of the non–class II exons compared in Table 1, only those of β_2-microglobulin align precisely end to end with those of E_α, employing the same split codon rule. The β_2-microglobulin gene also has the bulk of its 3′ untranslated sequence isolated as a distinct exon some distance 3′ to the last coding sequence (1.1 kb for β_2-microglobulin, 0.8 kb for DR_α).

Though Table 1 suggests that the domains compared diverged from a common anecestor, the evolutionary relationships between the entire genes is unclear. Non-α2-like sequences might have been under much different selective constraints and simply have diverged beyond recognition. It is also conceivable that the α2-like domain has been placed in different genomic contexts through the evolutionary process of exon shuffling (*24*–*27*). The organizational similarities between E_α and β_2-microglobulin suggest a more direct evolutionary relationship. Regardless, there appear to be fundamental evolutionary relationships among the three classes of genes that regulate and mediate immune responsiveness—Ia antigens, transplantation antigens, and immunoglobulins.

JANET MCNICHOLAS
Department of Biological Sciences, Stanford University, Stanford, California 94305
MICHAEL STEINMETZ
TIM HUNKAPILLER
Division of Biology, California Institute of Technology, Pasadena 91125
PATRICIA JONES
Department of Biological Sciences, Stanford University
LEROY HOOD
Division of Biology, California Institute of Technology

References and Notes

1. J. Klein, *Biology of the Mouse Histocompatibility Complex* (Springer-Verlag, Berlin, 1975).
2. G. D. Snell, J. Dausset, S. Nathenson, *Histocompatibility* (Academic Press, New York, 1976).
3. J. A. Frelinger, J. Niederhuber, D. C. Shreffler, *Science* **188**, 268 (1975).
4. R. H. Schwartz, C. S. David, D. H. Sachs, W. E. Paul, *J. Immunol.* **117**, 531 (1976).
5. E. A. Lerner, L. A. Matis, C. A. Janeway, Jr., P. P. Jones, R. H. Schwartz, D. B. Murphy, *J. Exp. Med.* **152**, 1085 (1980).
6. J. M. McNicholas, D. B. Murphy, L. A. Matis, R. H. Schwartz, E. A. Lerner, C. A. Janeway, Jr., P. P. Jones, *ibid.* **155**, 490 (1982).
7. L. A. Matis *et al.*, *ibid.*, p. 508.
8. D. B. Murphy, in *The Role of the Major Histocompatibility Complex in Immunobiology*, M. E. Dorf, Ed. (Garland, New York, 1981), pp. 1–32.
9. J. L. Strominger *et al.*, in *ibid.*, pp. 115–172.
10. J. Kaufman and J. L. Strominger, *Nature (London)* **297**, 694 (1982).
11. H. Kratzin *et al.*, *Hoppe-Seyler's Z. Physiol. Chem.* **362**, 1665 (1981).
12. D. Larhammar, L. Schenning, K. Gustafsson, K. Wiman, L. Claesson, L. Rask, P. A. Peterson, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 3687 (1982).
13. D. Larhammar *et al.*, *Cell* **40**, 153 (1982).
14. M. Steinmetz *et al.*, *Nature (London)* **300**, 35 (1982).
15. A. M. Frischauf, H. Garoff, H. Lehrach, *Nucleic Acids Res.* **8**, 5541 (1980).
16. D. Fisher, M. Pecht, L. Hood, in preparation.
17. C. T. Wake, E. O. Long, M. Strubin, N. Gross, R. Accolla, S. Carrel, B. Mach, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
18. E. Sung, M. Hunkapiller, L. Hood, P. Jones, in preparation.
19. M. Steinmetz, K. W. Moore, J. G. Frelinger, B. T. Sher, F. W. Sher, E. A. Boyse, L. Hood, *Cell* **25**, 683 (1981).
20. P. A. Sharp, *ibid.* **23**, 643 (1981).
21. J. R. Parnes and J. G. Seidman, *ibid.* **29**, 661 (1982).
22. A. J. Korman, C. Auffray, A. Schamboeck, J. L. Strominger, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 6013 (1982).
23. M. Steinmetz *et al.*, *Cell* **24**, 125 (1981).
24. W. F. Doolittle, *Nature (London)* **271**, 581 (1978).
25. J. E. Darnell, *Science* **202**, 1257 (1978).
26. W. Gilbert, *Nature (London)* **271**, 501 (1978).
27. F. Crick, *Science* **204**, 264 (1979).
28. K. W. Moore, B. T. Sher, Y. H. Sun, K. A. Eakle, L. Hood, *ibid.* **215**, 679 (1982).
29. A. A. Reyes, M. Schold, K. Itakura, R. B. Wallace, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 3270 (1982).
30. M. Malissen, B. Malissen, B. R. Jordan, *ibid.*, p. 893.
31. T. Kawakami, T. Takahashi, T. Honjo, *Nucleic Acids Res.* **8**, 3933 (1980).
32. Y. Yamawaki-Kataoka, T. Kataoka, N. Takahashi, M. Obata, T. Honjo, *Nature (London)* **283**, 786 (1980).
33. P. Hieter, E. Max, J. Seidman, J. Maizel, P. Leder, *Cell* **22**, 197 (1980).
34. A. Bothwell, M. Paskind, M. Reth, T. Imanishi-Kari, K. Rajewsky, D. Baltimore, *Nature (London)* **298**, 380 (1982).
35. S. V. Suggs, R. B. Wallace, T. Hirose, E. H. Kawashima, K. Itakura, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 6613 (1981).
36. A. F. Williams and J. Gagnon, *Science* **216**, 696 (1982).
37. Supported by grants from the National Institutes of Health. We thank D. Fisher for his guidance throughout this work, especially for help with the deoxyribonuclease deletion subcloning and sequencing strategy. We thank A. Korman and J. Strominger, D. Larhammar, and P. Peterson for making the class II DNA sequences and preprints available to us prior to publication and for helpful discussions. We also thank K. Minard for preparation of sequencing gels and B. Larsh for preparation of the manuscript. J.M. is a postdoctoral fellow of the Arthritis Foundation. M.S. was supported by a Senior Lievre fellowship from the California Division of the American Cancer Society.

23 August 1982; revised 20 September 1982

Reprint Series
19 August 1983, Volume 221, pp. 750–754

SCIENCE

Nucleotide Sequence of a Light Chain Gene of the Mouse I-A Subregion: $A\beta^d$

Marie Malissen, Tim Hunkapiller, and Leroy Hood

Nucleotide Sequence of a Light Chain Gene of the Mouse I-A Subregion: $A\beta^d$

Abstract. *Ia (I region–associated) antigens are cell-surface glycoproteins involved in the regulation of immune responsiveness. They are composed of one heavy (α) and one light (β) polypeptide chain. We have sequenced the gene encoding the $A\beta^d$ chain of the BALB/c mouse. The presence of six exons is predicted by comparison with the complementary DNA sequences of human β chains and with partial protein sequence data for the $A\beta^d$ polypeptide. Sequence comparisons have been made to other proteins involved in immune responses and the consequent implications for the evolutionary relationships of these genes are discussed.*

The major histocompatibility complex (MHC) of the mouse is a cluster of genes encoding at least three different classes of proteins involved in immune responses (*1*). Class I molecules, the classic transplantation antigens, and class II molecules are cell-surface, membrane-bound glycoproteins. Class III molecules are serum protein components of the complement pathway. While class I proteins are found on essentially all cells, class II representatives appear limited mainly to the surface of B cells and to antigen-presenting cells such as macrophages. Class II proteins are required for antigen presentation and lymphocyte interactions involved in the activation and differentiation of antibody-producing cells or B cells (*2*). Murine class II genes are located in the I (immune response) region of the MHC and are consequently referred to as I region–associated or Ia antigens (*3*).

Two types of Ia antigen have been defined in mice, I-A and I-E. Both are composed of two noncovalently linked polypeptides, a heavy chain (α) of about 34,000 daltons and a light chain (β) of about 28,000 daltons (molecular size). In the past 2 years, advances in protein microsequencing and recombinant DNA technologies have led to a definition of the primary structure of various human and murine class II molecules (*4–14*). Both α and β polypeptides may be divided into two external domains of approxi-

mately 90 residues each (termed α_1 and α_2 for the α chain or β_1 and β_2 for the β chain), a transmembrane region about 30 residues long, and a cytoplasmic tail of about 15 residues.

We have determined the complete nucleotide sequence of the class II $A\beta^d$ gene. This structure, together with class I and class II genes that have been analyzed, permit us to discuss several interesting evolutionary features about the supergene family including the MHC and antibody genes.

An $A\beta$ gene was isolated from a cosmid library constructed from BALB/c (d haplotype) sperm DNA by screening with a complementary DNA (cDNA) probe of a human I-A region homolog, DC-β (*15*). We have sequenced 5.8 kilobases (kb) of cosmid clone 34.2 overlapping the region of DC-β hybridization. The sequencing strategy and gene organization are shown in Fig. 1 and the DNA sequence in Fig. 2. The predicted amino-terminal protein sequence is identical to that previously defined for the BALB/c $A\beta^d$ protein (*4*, *5*).

The coding sequence of $A\beta^d$ is distributed among six exons. All exons have the appropriate donor and acceptor RNA splice sites. The first exon encodes a 27-residue-long putative signal peptide with 25 consecutive hydrophobic or neutral amino acids. It is identifiable because it begins with a methionine and the first four amino acids of the mature protein. The length of the $A\beta^d$ messenger RNA, 1.4 kb, and the lack of hybridization of probe 1 (Fig. 1) with BALB/c poly(A)$^+$ (polyadenylated) messenger RNA (mRNA) (see below) limits the extent of the 5′ untranslated region to less than 150 bp upstream of the initiation codon. Consistent with this, the canonical promoter-associated sequence CCAAT (C, cytosine; A, adenine; T, thymine) is found about 135 bp 5′ to the methionine codon. However, no recognizable candidate for the Hogness-Goldberg TATA box can be found within the 200-bp stretch 5′ of this codon. The $E\alpha^k$ gene also lacks a canonical TATA box (*13*), although it does have an AT-rich area missing from $A\beta^d$. It remains to be seen whether the partial or complete absence of such sequences will be a common feature of murine class II genes and, if so, whether this will have any significance for their regulation.

The other five exons were defined on the basis of similarity with two human cDNA sequences: DC-β (*11*) and DR-β (*12*). The two external domains are encoded by exons 2 and 3. Each contains a pair of cysteines, presumably involved in intradomain disulfide bridges. The predicted β_1 domain amino acid sequence contains one potential glycosylation site (Asn-Glu-Thr) beginning at residue 19. This site is also conserved in both human sequences. The fourth exon encodes three different structural elements: a hydrophilic-connecting peptide (amino acids 190 to 199), a hydrophobic or uncharged transmembrane segment (200 to 220), and a hydrophilic cytoplasmic tail (221 to 226). The transmembrane element is apparently anchored in the cell membrane by three basic residues (Arg-His-Arg) on the cytoplasmic side. Exon 6 encodes the last four amino acids of the cytoplasmic tail. A poly(A) addition signal (AATAAA) is found 280 nucleotides 3′ to the termination codon. The predicted 3′ untranslated sequence is found by computer analysis to be weakly, but recognizably similar to that of the DC-β cDNA.

The cytoplasmic region of the DR-β cDNA is eight codons longer than that of the DC-β cDNA. Alignment of all three β sequences suggests that the size difference is internal and not simply a 3′ truncation of the DC-β coding sequence relative to that of the DR-β cDNA. This relative insertion is located between exon 4 and exon 6 and led us to suspect there may be an additional cytoplasmic exon in the intervening sequence. A search of this intron revealed only one 24-base sequence bounded by the appropriate RNA donor and acceptor splice signals with an open reading frame in phase with the adjacent exons that shares similarity with the extra sequence of the DR-β clone. A restriction fragment including this sequence (probe 3, Fig. 1) was hybridized to BALB/c spleen poly(A)$^+$ mRNA on a nitrocellulose filter (Northern blot) (*16*) yielding a band at 1.4 kb, which is the same size as the $A\beta^d$ mRNA. The same mRNA did hybridize to another coding sequence (probe 2) but did not hybridize to a noncoding region (probe 1). We therefore include exon 5 as part of the expressed mouse gene sequence and assume that the DC-β gene

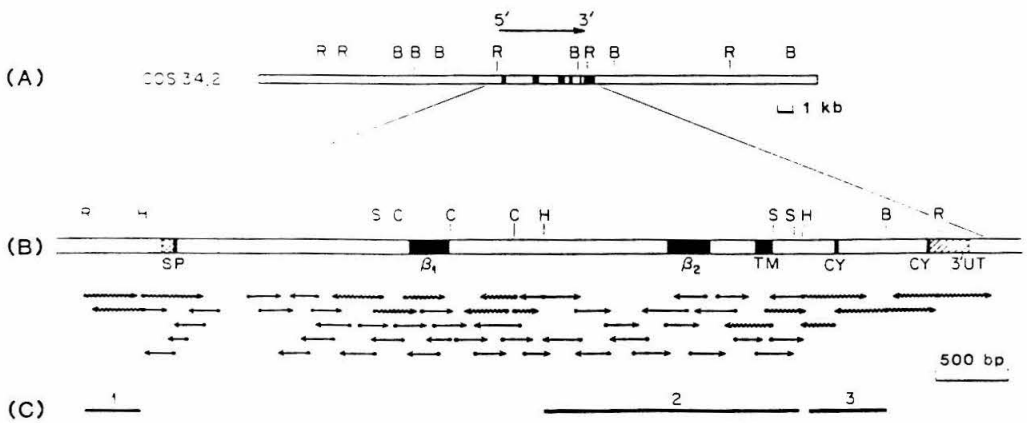
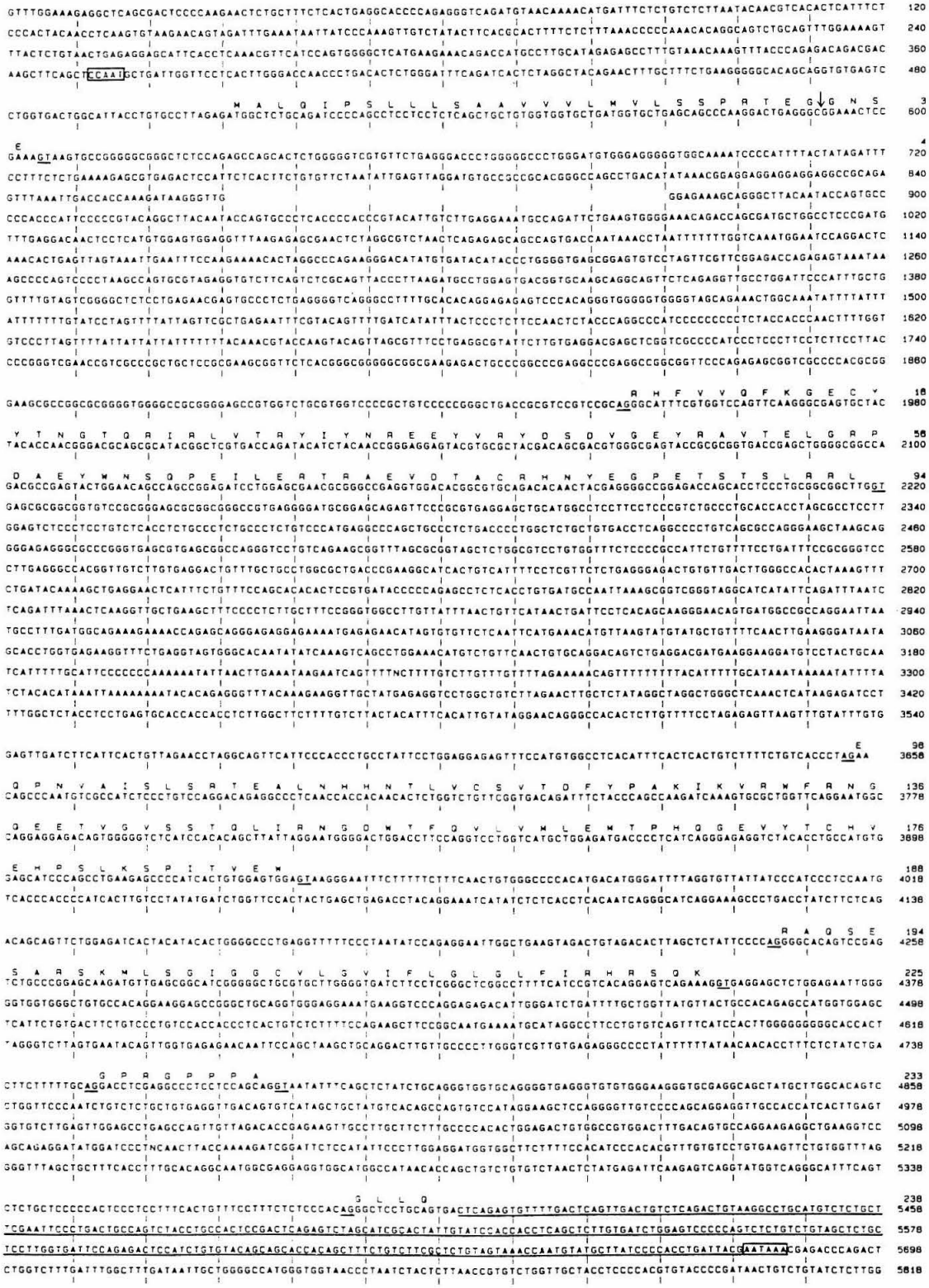


Fig. 1. The organization and sequencing strategy for the $A\beta^d$ gene. (A) Partial restriction map of the cosmid clone 34.2. Filled areas indicate exons. The arrow above shows the direction of transcription. Restriction enzyme names have been abbreviated as *R* (Eco RI) and *B* (Bam HI). (B) Structure and sequence strategy for the $A\beta^d$ gene. Arrows beneath indicate the sequencing strategy. Straight arrows give the sequence obtained by the dideoxy chain termination method (*20*). The two Hind III fragments were subcloned into the Sma I site of M13mp8 (*21*) in both orientations. Each subclone was used to generate a series of overlapping subclones by a deoxyribonuclease I deletion technique (*22*). Wavy arrows indicate the direction and length of sequences obtained by the chemical degradation method (*23*). *SP* refers to the signal peptide, β_1 and β_2 to the two external domains, *TM* the transmembrane portion of the protein, *CY* the cytoplasmic part, and *3′UT* the 3′ untranslated region of the mRNA. The region between the two dotted lines shown in the gene map has not been sequenced. Abbreviations for restriction enzymes not listed in (A) are *H* (Hind III), *S* (Sac I), and *C* (Sac II). (C) Different probes used in the Northern blot analysis (*16*).



has lost this exon subsequent to the divergence of humans and mice (*17*).

Similarity comparisons were made between defined regions of mouse and human Ia polypeptides. The greater similarity of $A\beta^d$ to DC-β (76 percent) than to DR-β (65 to 68 percent) supports the model that DC-β is the human homolog of the mouse I-Aβ molecule (*18*). The relative similarities between the equivalent domains of $A\beta^d$ and DC-β indicate that the first domain is the more divergent of the two external domains (66 and 86 percent, respectively). The high degree of similarity between the connecting peptides (90 percent) and transmembrane regions (86 percent) of human and mouse sequences is rather surprising in view of their presumably general functions. Conservation between similar regions of I-E–like α chains, class I molecules, and immunoglobulins is much less. This level of conservation leads us to suspect that the connecting and transmembrane regions may be involved in the interaction with the α chain or some other membrane protein rather than serving just as hydrophobic anchorage. The size difference between the $A\beta^d$ and DC-β cytoplasmic regions implies that significant variation of this region may have less severe selective consequences than it would for other regions of the gene.

Computer analysis reveals no significant similarity between the β_1 and β_2 domains of $A\beta^d$ nor between β_1 and any domain of any other gene involved in immune responses. However, the β_2 domain is significantly similar to the α_2 domains of class II proteins, the third external domain (α_3) of class I heavy chains, β_2-microglobulin, and immunoglobulin constant-region domains. A computer-aided alignment was made between most of the known representatives of these "Ig-like" domains. The alignment (data not shown) shows a striking conservation of functionally similar residues responsible in immunoglobulins for the β-pleated sheet regions of a tertiary structure referred to as the antibody fold (*19*). The alignment shows that the class I, class II, and β_2-microglobulin sequences are roughly comparable in their similarities to each other as well as comparably less similar to immunoglobulin sequences, establishing two evolutionary subgroups.

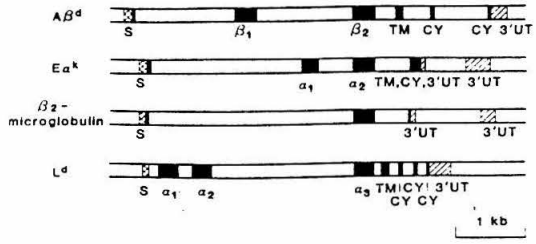


Fig. 3. Comparison of the intron-exon organization of class I, class II, and β_2-microglobulin (*24*). The $A\beta^d$ is representative of a class II β chain, the $E\alpha^k$ (*13*) of a class II α chain, and L^d of a transplantation antigen heavy chain (*25*). Solid, filled areas encode residues present in the mature proteins, crossed areas show the portions encoding the signal peptides, and hatched boxes represent the 3′ untranslated sequences. The abbreviations are the same as in Fig. 1; β_1 and β_2 and α_1 and α_2 refer to exons encoding the external domains of the protein.

The organizations of the $A\beta^d$ gene, a class II α gene, a class I heavy chain, and β_2-microglobulin are shown in Fig. 3. The most striking comparison to be made from Fig. 3 is between class II α genes and β_2-microglobulin. Note particularly (i) the distances between the signal peptide and the Ig-like domain, (ii) that the first few amino acids of each mature protein are encoded with the signal peptide by the first exon, and (iii) the rare occurrence of an intervening sequence in the 3′ nontranslated region of both genes. The $A\beta^d$ gene shares the first two features but not the third one. Also, like the class I genes and not the class II α genes, the cytoplasmic coding sequence of $A\beta^d$ occurs on multiple exons.

These organizational comparisons support the suggestion (*14*) that class II α chains and β_2-microglobulin are more directly related evolutionarily than are class II α and β chains. It also was suggested that class I and class II genes might not be paralogous evolutionarily, but are each the result of de novo gene construction through the process of "exon shuffling." However, the structure of the $A\beta^d$ gene as presented here is not as dissimilar to class I genes as is the Eα gene. Moreover, though not significantly similar over their entire length, the first external domains of class I heavy chains (α_1) and the β_1 domains of class II β chains do share a highly conserved peptide sequence about residue 40 (Val-Arg-Phe-Asp-Ser-Asp). $A\beta^d$ differs by one conservative substitution (Phe to Tyr). The shorter α_1 domains of class II α chains possess a similar and completely conserved sequence, Phe-Asp-Gly-Asp, about residue 30, but have no better than random similarity beyond this with the α_1 domains of class I or β_1 domains of class II. The conservation of sequence across all examples of class I and class II molecules and the structural similarities of the genes support a direct evolutionary connection between class I and class II genes, particularly those of class II β chains.

Antigen triggering of cellular and humoral immune responses requires the recognition by different T-cell populations of antigens presented in the context of class I or class II molecules. The charged nature of the conserved peptide probably ensures its residence on the exterior of the molecule. As a conserved, exposed determinant, this peptide sequence may play a homologous role in T cell recognition of both class I and class II antigens. It is possible then that not only are class I, class II, and β_2-microglobulin related by distant ancestry to immunoglobulin, but that class I and class II antigens are derived from the same ancestral two-chain polypeptide involved in prototypical lymphocyte interactions.

Note added in proof: Comparison of the $A\beta^d$ sequence with that of the b-haplotype allele (*26*) indicates a relative deletion of an Alu-like (β_1) repeat sequence in the second intron of $A\beta^d$. Computer analysis reveals a repeat-like element only vaguely similar to β_1 about 300 bp 5′ of the β_2 domains of both alleles.

Fig. 2. The DNA sequence of 5.8 kb of the cosmid clone 34.2 containing the Aβ gene. The RNA splicing donor and acceptor sites (GT-AG) (G, guanine) are underlined as is the probable 3′ untranslated region. A possible CCAAT box and poly(A) addition signal sequence are boxed. The gap shown between bases 870 and 871 represents less than 200 bases of undetermined sequence. The predicted amino acid sequences encoded by the exons are shown above the DNA. Single-letter amino acid codes are placed above the middle bases of the translated codons. The single-letter amino acid code is abbreviated as follows: A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; and W, tryptophan.

MARIE MALISSEN
TIM HUNKAPILLER
LEROY HOOD

Division of Biology,
California Institute of Technology,
Pasadena 91125

References and Notes

1. J. Klein, A. Juretic, C. Baxevanis, Z. Nagy, *Nature (London)* **291**, 455 (1981).
2. A. Singer, Y. Asano, M. Shigeta, K. S. Hathcock, A. Ahmed, C. G. Fathman, R. J. Hodes, *Immunol. Rev.* **64**, 137 (1982).
3. H. O. McDevitt, *J. Immunogenet.* **8**, 287 (1981).

4. J. M. Cecka, M. McMillan, D. B. Murphy, H. O. McDevitt, L. Hood, *Eur. J. Immunol.* **9**, 955 (1979).
5. E. Sung, M. W. Hunkapiller, L. E. Hood, P. P. Jones, in preparation.
6. H. Kratzin *et al.*, *Hoppe-Seyler's Z. Physiol. Chem.* **362**, 1665 (1981).
7. C. Auffray, A. J. Korman, M. Roux-Dosseto, R. Bono, J. L. Strominger, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 6337 (1982).
8. C. O. Benoist, D. J. Mathis, M. R. Kanter, V. E. Williams II, H. O. McDevitt, *ibid.* **80**, 534 (1983).
9. A. J. Korman, C. Auffray, A. Schamboeck, J. L. Strominger, *ibid.* **79**, 6013 (1982).
10. J. S. Lee *et al.*, *Nature (London)* **299**, 750 (1982).
11. D. Larhammar *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 3687 (1982).
12. E. O. Long, C. T. Wake, J. Gorski, B. Mach, *EMBO J.* **2**, 389 (1983).
13. D. J. Mathis, C. O. Benoist, V. G. Williams II, M. R. Kanter, H. O. McDevitt, *Cell* **32**, 745 (1983).
14. J. McNicholas, M. Steinmetz, T. Hunkapiller, P. Jones, L. Hood, *Science* **218**, 1229 (1982).
15. M. Steinmetz *et al.*, *Nature (London)* **300**, 35 (1982).
16. M. Kronenberg, M. M. Davis, P. W. Early, L. E. Hood, J. D. Watson, *J. Exp. Med.* **152**, 1745 (1980).
17. The fifth exon sequence has been demonstrated in the cDNA sequence of the d-hapiotype allele of Aβ [E. Choi, K. McIntyre, R. N. Germain, J. G. Seidman, *Science* **221**, 283 (1983)].
18. J. R. Bono and J. L. Strominger, *Nature (London)* **299**, 836 (1982).
19. D. Beale and A. Feinstein, *Quant. Rev. Biophys.* **9**, 135 (1976).
20. F. Sanger, A. R. Coulson, B. G. Barrel, A. J. H. Smith, B. A. Roe, *J. Mol. Biol.* **143**, 161 (1980).
21. J. Messing and J. Vieira, *Gene* **19**, 269 (1982).
22. S. Anderson, *Nucleic Acids Res.* **9**, 3015 (1981).
23. A. M. Maxam and W. Gilbert, *Methods Enzymol.* **65**, 499 (1980).
24. J. R. Parnes and J. G. Seidman, *Cell* **29**, 661 (1982).
25. K. W. Moore, B. T. Sher, Y. H. Sun, K. A. Eakle, L. Hood, *Science* **215**, 679 (1982).
26. D. Larhammar *et al.*, *Cell*, in press.
27. We thank M. Steinmetz for his help during the initiation of this work, K. Minard for preparation of the sequencing gels, and B. Larsh and S. Olive for preparation of the manuscript. Supported in part by NIH grants (L.H.) and by the Fondation pour la Recherche Médicale (M.M.).

16 June 1983

COMPARISON OF EXON 5 SEQUENCES FROM 35 CLASS I GENES OF THE BALB/c MOUSE

By KURT A. BRORSON, STEPHEN W. HUNT III, TIM HUNKAPILLER, Y. HENRY SUN, HILDE CHEROUTRE, DEBORAH A. NICKERSON, AND LEROY HOOD

From the Division of Biology, California Institute of Technology, Pasadena, California 91125

The mouse class I MHC molecules are structurally related 45-kD cell surface glycoproteins that associate noncovalently with β_2-microglobulin, a 12-kD polypeptide (1). Class I molecules can be divided into two groups on the basis of their pattern of expression and their function. The transplantation antigens, H-2K, H-2D, and H-2L, are expressed on most somatic cells and present viral antigens to CTLs (2). The other group, the nonclassical class I molecules, exhibit a generally more restricted tissue distribution and are probably not involved in antigen presentation (3–7).

The BALB/c mouse has at least 35 class I genes that map to five genetic loci: *K*, *D*, *Qa*, *Tla*, and *Hmt* (8–11; Fig. 1). The classical transplantation antigen genes, K^d, D^d, and L^d map to the *K* and *D* loci, as do four other class I genes: $K2^d$, $D2^d$, $D3^d$, and $D4^d$ (12, 13). The *Qa* and *Tla* loci together contain 28 known class I genes, including some shown to encode nonclassical class I molecules (6, 14–16). In BALB/c mice, the eight *Qa* region genes are named $Q1^d$, $Q2^d$, $Q4^d$, $Q5^d$, $Q6^d$, $Q7^d$, $Q8/9^d$, and $Q10^d$, and the 19 *Tla* region genes are named $T1^c$ through $T18^c$ and 37^c. The newly described *Hmt* region contains at least three class I genes (10), including the *Thy-19.4* gene (11), which is included in this study.

Class I genes contain 6–8 exons (14, 17, 18). Exon 1 encodes a hydrophobic leader segment that is proposed to assist in the transport of the molecule to the cell surface and is cleaved post-translationally (19). Exons 2, 3, and 4 each encode the three 90-amino acid external domains: α1, α2, and α3. A short external connecting peptide, as well as the transmembrane domain and part of the cytoplasmic segment that includes charged anchoring residues, are encoded by exon 5 (Fig. 2). Exons 6, 7, and 8 encode the remainder of the cytoplasmic domain. Analysis of exon 5 sequences shows that they are generally not conserved for direct sequence similarity, but rather for maintaining hydrophobicity in the transmembrane stretch that they encode (20). Certain class I gene products, like those of the $Q4^d$ and $Q10^d$ genes,

This work was supported by National Institute of Health Grant AI-17565. S. W. Hunt is a Special Fellow of the Leukemia Society of America. D. Nickerson was a visiting associate from the University of South Florida, Tampa, FL. S. W. Hunt's present address is the Division of Rheumatology and Immunology, The University of North Carolina, Chapel Hill, NC 27599. Y. H. Sun's present address is the Institute of Molecular Biology, Academia Sinica, Nankang, Taipei 11529, Taiwan, Republic of China. H. Cheroutre's present address is the Department of Microbiology and Immunology, University of California, Los Angeles, CA 90024

Address correspondence to K. A. Brorson, Division of Biology-147-75, California Institute of Technology, Pasadena, CA 91125

are secreted and do not maintain a hydrophobic transmembrane domain. The $Q7^d$ gene product, the Qa-2 antigen, has a typical hydrophobic transmembrane domain and charged anchor residues, yet is linked to the cell surface via a phosphatidylinositol linkage (15). The transmembrane domain of the Qa-2 molecule is proposed to be cleaved before expression on the cell surface.

This report compares the exon 5 DNA sequences of the 35 known class I genes of the BALB/c mouse. Such a comparison can reveal which of these exons can encode a hydrophobic transmembrane, and whether the putative gene product could be membrane bound or secreted. Whereas the structure of the external α1, α2, and α3 domains has been resolved for at least one human class I antigen (21), no direct structural data exists for the transmembrane domains for the class I molecules. Therefore, an analysis of the predicted amino acid sequences of these genes could reveal what amino acid sequence and structural considerations are important for the function of the transmembrane domains. Analysis of the sequences reveals that, in spite of extensive nucleotide sequence variation, only four class I gene fifth exons, those from the $Q10^d$, $T5^c$, $T11^c$, and $T12^k$ genes, have frame shifts or stop codons that terminate their translation and prevent them from encoding a domain that is hydrophobic and long enough to span a lipid bilayer. Of the remaining fifth exons, 27 can encode membrane-spanning domains that resemble those of the classical transplantation antigens in that they can be divided into a proline-rich connecting peptide, a transmembrane segment, and a cytoplasmic segment with anchoring basic residues. In addition, hydrophobic moment analysis of the predicted transmembrane domains reveals that several, including those of the Qa-2 and TL antigens, are sufficiently amphipathic to promote intramembrane protein interactions. The conservation of the ability to encode a potentially functional transmembrane domain in the majority of the fifth exons suggests that selective pressure exists on them to remain functional, possibly because the majority of class I genes, including the divergent ones, are functionally important.

Materials and Methods

Sequencing of Transmembrane Exons. Individual class I genes or gene fragments were cloned from BALB/c MHC class I cosmids (8) into M13mp18- or pUC18-derived vectors. DNA sequencing was performed using the dideoxynucleotide chain termination method (22). Sequencing was primed with an oligonucleotide (5' ACCTTCCAGAAGTGGGCA 3') derived from a conserved area of the fourth exon of the L^d gene (23). This primer was chosen because the same sequence occurs in the fourth exon of several divergent class I genes, including the $H\text{-}2K^d$, D^d, L^d, $T13^c$ and $Q7^d$ (24) genes, and hence, is presumably highly conserved in most class I genes. Since exon 5 is generally ~120 nucleotides long and 210 nucleotides downstream of the primer, it was possible to determine the complete exon 5 sequence of all unpublished genes on one strand with one set of sequencing reactions. Exon 5 sequences that could not be directly aligned with previously reported sequences were also sequenced on the opposite strand using complementary oligonucleotide primers derived from intron 5 sequence. The fifth exons that can be aligned and were not sequenced on two strands were those from the $K2^d$, $Q1^d$, $Q4^d$, $Q5^d$, $Q6^d$, $Q10^d$, and $T3^c$ genes.

Sequence Alignments and Comparisons. Sequence alignments were performed using the method of Needleman and Wunsch (25), which inserts gaps into one or the other of the sequences in a pairwise comparison to maximize the similarity between the two sequences. In the percentage sequence similarity calculation, a gap of any size is counted as one mismatch, whereas unmatched sequences at either end are not counted.

After alignment, pairs of sequences were analyzed at each position for possible and observed silent and replacement substitutions (26). A single base change that does not change the predicted translation of coding region sequence is considered to be silent, while one that does change the predicted translation is considered to be a replacement. Each possible pairing of aligned sequences was analyzed, and substitutions were totaled for each category.

Analysis of Translated Exon 5 Sequences. The translated exon 5 sequences were analyzed by an algorithm that calculates the hydrophobicity of 21-amino acid stretches of the sequence (27). The hydrophobicity values of individual amino acids are taken from a consensus scale adapted from five separate hydrophobicity measurements (28). The method calculates which 21-amino acid stretch has the highest hydrophobicity value, thereby predicting which segment, if any, best defines the transmembrane domain.

Within the predicted 21-amino acid transmembrane segment, the hydrophobic moments, a measure of amphipathicity, of 11-amino acid stretches were calculated using the equation of Eisenberg et al. (27). The highest hydrophobic moment value for each transmembrane was plotted against the hydrophobicity value for the corresponding 11-amino acid stretch. The empirically defined area of the graph in which the point falls predicts where the predicted helix is likely to be found relative to the membrane, and whether it resides in the membrane alone or in association with another protein.

Results and Discussion

Exon 5 Sequences and Groups. The DNA sequences of the fifth exons of 35 BALB/c class I genes are shown in Fig. 3. In most cases, intron 4 and a portion of intron 5 are also included. The sequences were obtained from subclones of the BALB/c cosmids in this study, or from published sequencies. In most cases, exon 5 is identified by nucleotide similarity to known fifth exons, while in the cases of the *T7*, *T15*, and *Thy-19.4* genes, exon 5 is identified by the hydrophobicity of the translated amino acids, and relative position 3′ of exon 4. Exon boundaries are identified by comparison to class I genes for which spliced cDNA clones have been isolated (29, 30, Hunt, S., K. Brorson, H. Cheroutre, and L. Hood, manuscript in preparation), and by position of consensus splice sites. Donor splice sequences are not found in the *T4*, *T5*, *T7*, and *T15* fifth exons. The fifth exons of the *T7* and *T15* genes are interrupted by a B1 short interspersed repetitive element (31) after 143 bp, while those of the *T4* and *T5* genes are similar to other fifth exons for the first 46 and 58 bp, respectively, but contain nonhomologous sequences beyond what appears to be a

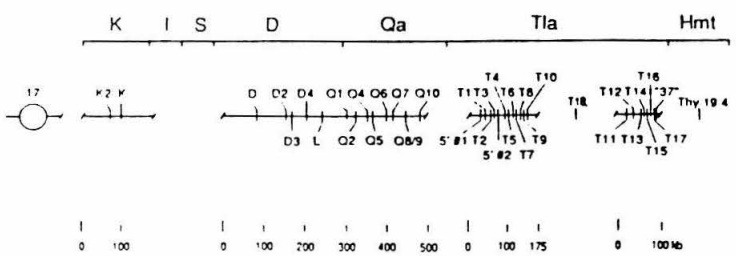


FIGURE 1. Map of class I genes in the BALB/c MHC. Class I genes map to the *K*, *D*, *Qa*, *Tla*, and *Hmt* regions. The *I* and *S* regions contain class II and complement genes, respectively. The order of the *Tla* region gene clusters is unknown, as is the distance between the *K*, *D*, *Tla*, and *Hmt* regions. The upper line represents the genetic map and the gene clusters are indicated below.

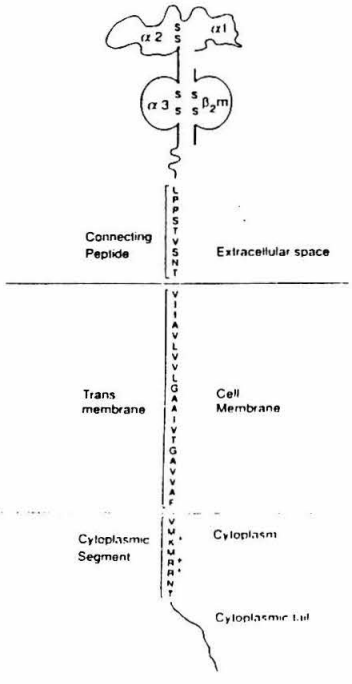


FIGURE 2. Model of typical class I transmembrane domain, H-2K^d. Connecting, transmembrane, and cytoplasmic segments are shown in which environment they are predicted to reside by the criteria used in this report (27).

recombinational or gene conversion boundary. The predicted reading frames are identified by the hydrophobicity of the translated amino acids, and by conformity to the reading frame established by the fourth exon.

The fifth exon sequences are assigned to the same group if they share at least 75% similarity with each other (32). The exon five sequences can be assigned to seven nonoverlapping groups. The largest group includes all of the *H-2* and *Qa* loci genes, and in addition, several even numbered *Tla* region genes: *T4*^{c}, *T6*^{c}, *T8*^{c}, *T10*^{c}, *T14*^{c}, *T16*^{c}, and *37*^{c}. A second group includes the *T1*^{c}, *T3*^{c}, *T11*^{c}, and *T13*^{c} genes, while a third includes the *T2*^{c}, *T5*^{c}, and *T12*^{c} genes. Finally, the *T9*^{c}/*T17*^{c} and *T7*^{c}/*T15*^{c} gene pairs form two additional groups, while the *T18*^{c} and *Thy-19.4* genes form two additional single gene groups. Consensus sequences are derived for each group based on the most frequent nucleotide used at each position.

Nucleotide sequence similarity among members of each group ranges from 73 to 99% (Table I). No two members of any group are exactly identical, making the exon 5 sequences diagnostic for the identification of BALB/c class I genes. Among members of each group, several types of mutational events have occurred subsequent to the duplications that created them, including nucleotide substitutions and short deletions. In addition, the exon 5 of the *Q1*^{d} gene has an extra 18 bp that matches 15 of the 18 nucleotides immediately following it, and thus, it is probably the product of an internal sequence duplication. Interestingly, an 18-bp insertion that matches

TABLE I

Percent Similarity of Fifth Exons within Groups

		Similarity of exons																			
Group	Exons	K^a	$K2^a$	D^d	$D2^d$	$D3^d$	$D4^a$	L^d	$Q1^a$	$Q2^a$	$Q4^a$	$Q5^a$	$Q6^a$	$Q7^a$	$Q8/9^d$	$Q10^d$	$T4^c$	$T6^c$	$T8^c$	$T10^c$	$T14^c$	$T16^c$
1	$K2^d$	86																			
	D^d	82	84																		
	$D2^d$	82	85	82																	
	$D3^d$	85	85	85	87																
	$D4^a$	87	90	85	85	85															
	L^d	82	85	85	85	88	85														
	$Q1^d$	85	85	85	85	91	87	85													
	$Q2^d$	89	89	86	88	91	90	90	91												
	$Q4^d$	84	86	78	85	87	86	88	85	88											
	$Q5^d$	83	84	98	83	86	85	85	85	87	79										
	$Q6^d$	83	84	78	84	86	86	87	83	87	97	78									
	$Q7^a$	83	86	82	88	90	86	87	88	90	87	83	86								
	$Q8/9^d$	83	85	81	87	89	86	86	87	84	86	82	87	99							
	$Q10^d$	73	78	76	82	79	79	84	78	8	80	77	79	81	80						
	$T4^c$	89	89	78	80	87	87	85	89	93	85	78	83	83	80	76					
	$T6^c$	81	84	81	81	83	86	83	88	88	81	82	80	81	80	75	93				
	$T8^c$	86	87	86	83	86	86	86	85	89	81	86	81	83	82	79	87	85			
	$T10^c$	83	85	82	85	85	83	83	87	89	80	83	79	84	83	76	85	81	82		
	$T14^c$	82	85	82	82	84	87	85	88	89	84	83	83	82	81	76	93	95	86	82	
	$T16^c$	87	88	87	84	87	87	87	86	90	82	87	81	84	83	80	89	86	99	83	87
	$3T$	88	88	87	86	92	89	87	92	94	89	88	88	88	87	79	93	90	89	88	91
		$T1^c$	$T3^c$	$T11^c$																	
2	$T3^c$	90																			
	$T11^c$	94	93																		
	$T13^c$	91	98	95																	
		$T2^c$	$T5^c$																		
3	$T5^c$	95																			
	$T12^c$	98	91																		
4	$T7^c$ vs. $T15^c$	98																			
5	$T9^c$ vs. $T17^c$	99																			

In the comparisons, only the first 58 bp of the $T5^c$ fifth exon and 46 bp of the $T4^c$ fifth exon are used. The full lengths of the fifth exons are compared between all of the other genes.

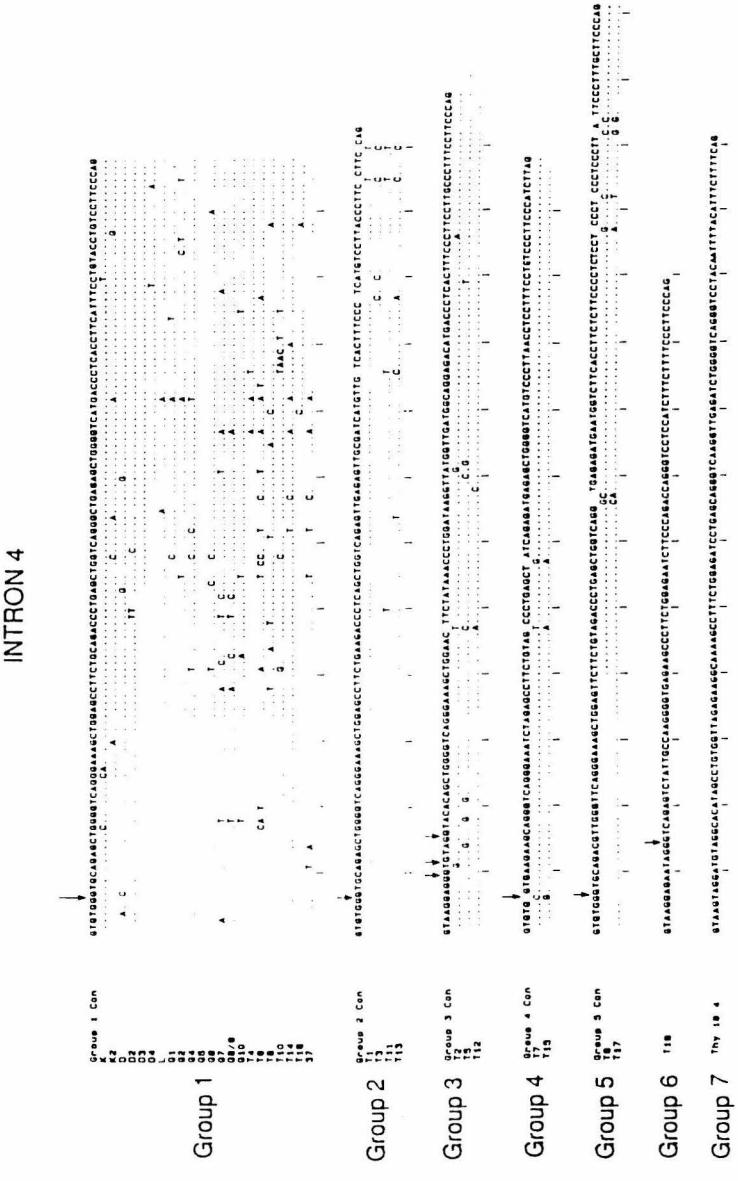
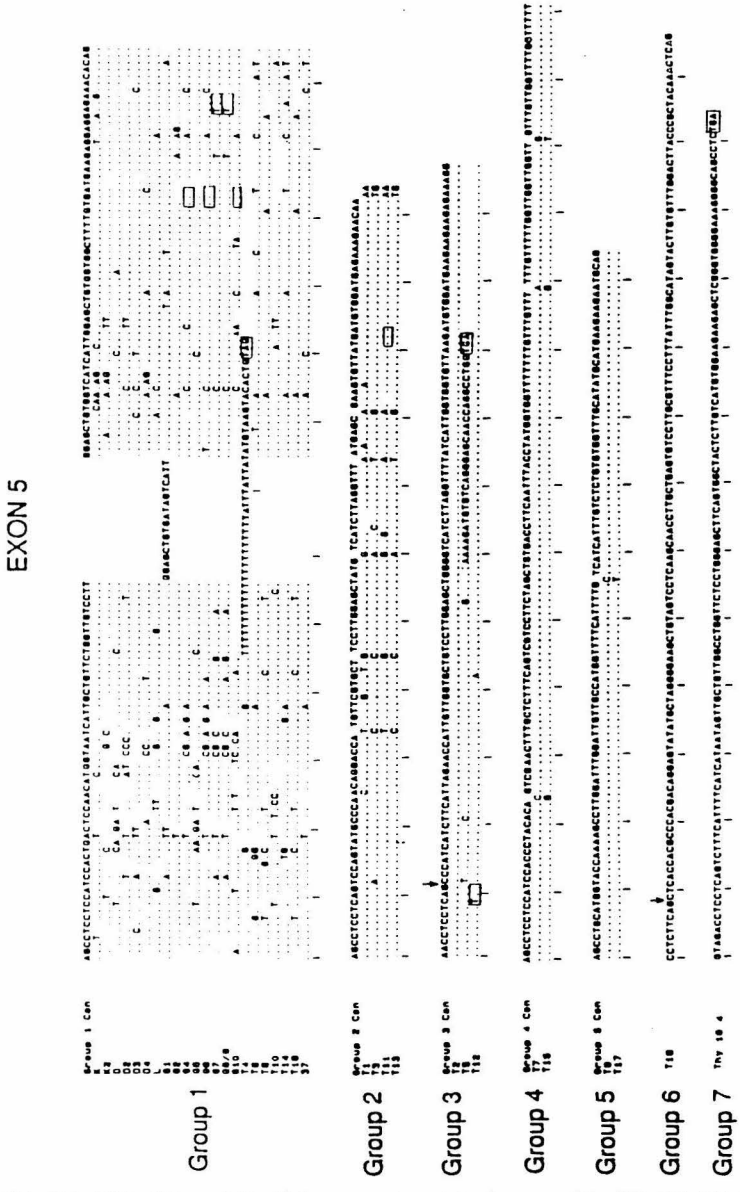


FIGURE 3 The DNA sequence of the fifth exons of the 35 BALB/c class I genes. Genes that were sequenced on one strand are the *K2*^{d}, *Q1*^{d}, *Q4*^{d}, *Q5*^{d}, *Q10*^{d}, and *T3*^{c} genes. All others were sequenced on both strands or have been previously published: *L*^{d} (22); *K*^{d} (17); *T13*^{c} (13); *D*^{d} (59); *37* (60); *Thy-19.4*^{c} (11); *T1*^{c} (61); *D2*^{d} (62); *Q7*^{d}, *T9*^{c}, *T17*^{c}, *T18*^{c}, *D3*^{d}, *D4*^{d} (Hunt et al., manuscript in preparation); *Q5*^{d}, *Q6*^{d}, and *Q8/9*^{d} (I. Stroynowski, personal communication); *Q10*^{d} (N. Ulker, personal communication). Intron 4 and 5 sequences are included in most cases. In-frame translation termination codons found in the fifth exons of the *Q4*^{d}, *Q6*^{d}, *Q7*^{d}, *Q8/9*^{d}, *Q10*^{d}, *T4*^{c},



T5', *T7'*, *T11'*, *T12'*, *T15'*, and *Thy-19.4* genes are boxed. In the cases of the *T4'*, *T5'*, *T7'*, and *T15'* genes, the separation between exon 5 and intron 5 is arbitrary. In the *T4'* and *T5'* genes, the end of exon 5 is defined as the in-frame stop codon where translation terminates, while in the *T7'* and *T15'* genes, the B1 repeat element is arbitrarily included with the other intron 5 sequences. Possible alternative splice signals are indicated by arrows above the consensus sequences. These sequence data have been submitted to the EMBL/GenBank Data Libraries under the accession numbers X16197 through X16223. Figure continued on following page.

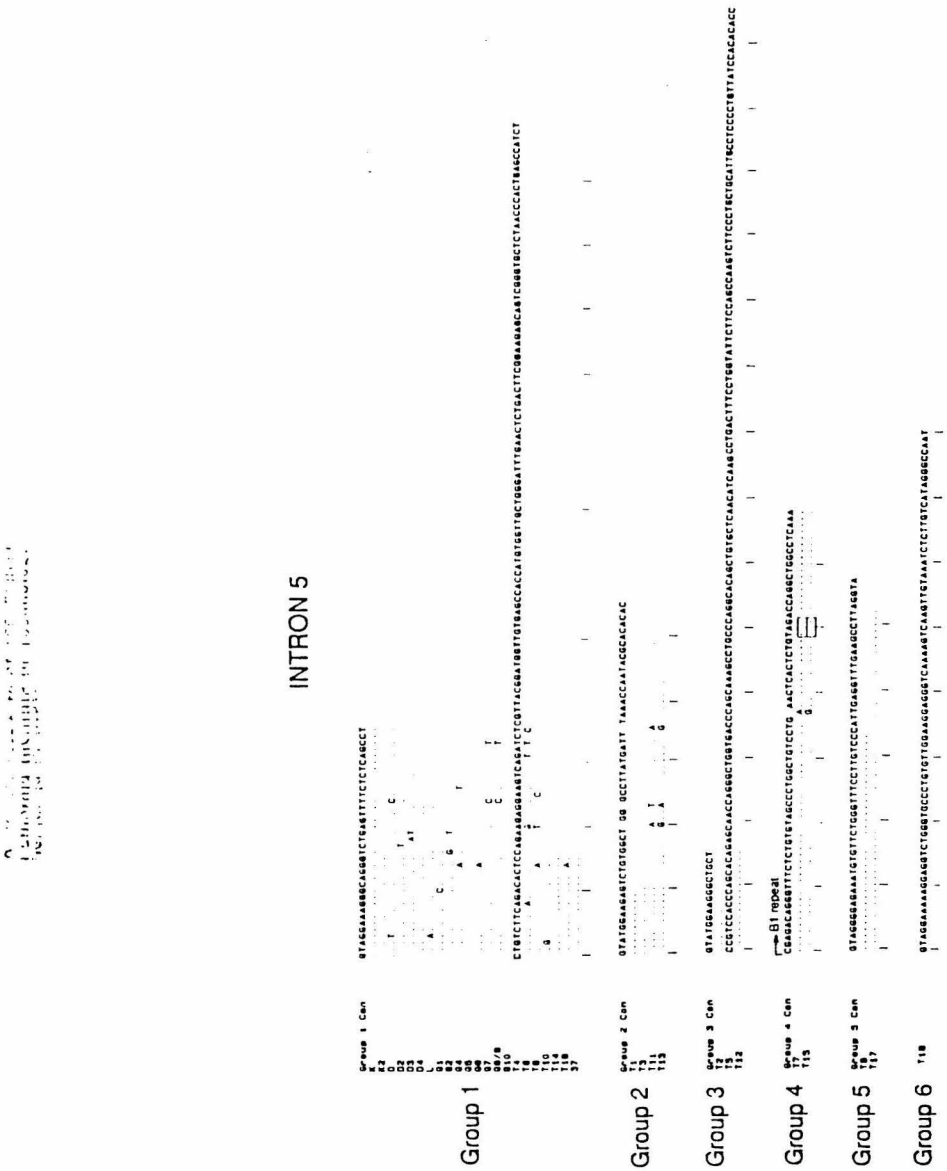


FIGURE 3. Continued. Legend on previous page.

the $Q1^d$ insertion in 13 of 18 nucleotides is also found in the same position in a rat class I gene (33; J. Howard, personal communication). Since the inserted sequence is about as similar to the rat sequence as it is to the 18-bp sequence following it, it is possible that this duplication event occurred before mouse/rat divergence. Alternatively, since this appears to be a single mutational event in both species, it is possible that the duplications occurred independently. Since the insertion in the $Q1^d$ exon 5 does not match precisely either the sequences immediately following it, or the homologous insertion in the rat fifth exon, it is unclear which of these two possibilities is the case.

Comparison of group consensus sequences reveals that some groups are related while others appear not to be (Table II). Groups 2 and 3 are ~73% similar to each other. Their members are different enough from each other to be classified as distinct groups based on the criteria of this report, but they are clearly evolutionarily related. The other groups are possibly related to each other since some pairings share as much as 54% similarity. Unlike the similarity between groups 2 and 3, it is unclear if 32–54% sequence similarity between these groups is the result of divergent evolution from an ancestral exon 5, or rather of convergent evolution of unrelated transmembrane exons. Codon usage that is restricted to maintain hydrophobicity of the translation could result in unrelated sequences attaining greater than random similarity. Thus, it is conceivable that the exon 5 sequences have multiple origins as the result of exon shuffling or de novo generation, and share similarity because of convergent evolution. This is most possible for the $T18^c$ fifth exon since it shares only 32–43% similarity to all of the other groups.

The existence of variation in the transmembrane exons in the class I gene family argues that their most important sequence consideration is the retention of a hydrophobic translation (20, 34). Extensive variation occurs in transmembrane domains when their only function is to anchor a protein to a membrane. To test whether selective pressure exists for the fifth exon sequences to retain their translation, synonymous and nonsynonymous mutation frequencies were determined for the group 1, 2, and 3 fifth exons (Table III). Since the members of groups 2 and 3 can be aligned,

TABLE II

Percent Similarity between Consensus Sequences of Fifth Exon Groups

	Similarity of exons					
Group	G1	G2	G3	G4	G5	G6
			%			
2	54					
3	53	73				
4	50	43	43			
5	41	38	48	46		
6	32	41	33	36	43	
7	48	38	49	40	42	37

Before the percent similarity calculation, the fifth exon sequences were aligned with gaps to maximize the result. Groups 1–7 consensus sequences are abbreviated as G1 through G7.

TABLE III

Replacement and Silent Site Mutation Frequencies in Exon 5 and Intron 4

	Replacement	Silent	Possible	Observed
Exon 5 Group 1	2,766/17,249 (16)	765/6,133 (12)	2.81	3.62
Exon 5 Groups 2 and 3 (aligned)	202/1,259 (16)	40/361 (11)	3.49	5.05
Intron 4 (Exon 4 read-through) Group 1	1,264/15,319 (8)	594/5,951 (10)	2.57	2.13
Intron 4 Groups 2 and 3 (aligned)	217/1,568 (14)	68/496 (14)	3.16	3.19

Frequencies are expressed both as a fraction of observed changes over possible changes, and as a percentage (in parentheses). Possible and observed replacement/silent ratios are also shown. The *T4*ᵃ and *T5*ᵃ fifth exons were omitted from this analysis since their 3′ portions were created by recombination events, not duplication and point mutation.

they were pooled to maximize the number of sites tested. As a contrast, synonymous and nonsynonymous mutation frequencies were also determined in the exon 4 read-through frame of intron 4. If selective pressure is exerted on a coding region sequence, the frequency of silent mutations is predicted to be higher than that of replacement mutations. As expected, silent and replacement mutations are approximately equivalent in the intron 4 sequences. However, in the fifth exons of groups 1, 2, and 3, replacement mutations have a higher frequency than silent mutations. This suggests that there is little selective pressure to maintain their protein encoding sequences other than for hydrophobicity, although the variation in the putative *T1a* region gene-encoded transmembranes may have evolved because they perform specialized functions that are different than those of the transplantation antigens.

In contrast to the fifth exon sequences, almost all exon 4 sequences are at least 80% similar to each other (Hunt, S., K. Brorson, H. Cheroutre, and L. Hood, manuscript in preparation, K. Brorson unpublished observations), supporting the concept that all of the class I genes evolved from a common ancester (35). It is interesting that the highly conserved fourth exons and the highly divergent fifth exons are separated only by a 120-nucleotide intron. Dot matrix identity plots between group consensus sequences (Fig. 4) reveal that, between groups, intron 4 is more conserved than exon 5, and that there are two general areas of conservation. One area is the splice acceptor site and the first ~10 bp of exon 5. The other area is the 5′ portion of intron 4, adjacent to the conserved exon 4. It is conserved among all of the genes except the *T18*ᵃ gene, and the *Thy-19.4* gene, where only the middle of the intron is conserved. Intron 4 is also more conserved than exon 5 when compared among groups (Table III). However, it is unlikely that this reflects selective pressure for their conservation, as would occur if read-through translation from exon 4 is important since the silent mutation frequency of the read-through frame is approximately equal to the replacement frequency in group 1, as well as in groups 2 and 3. Instead, the

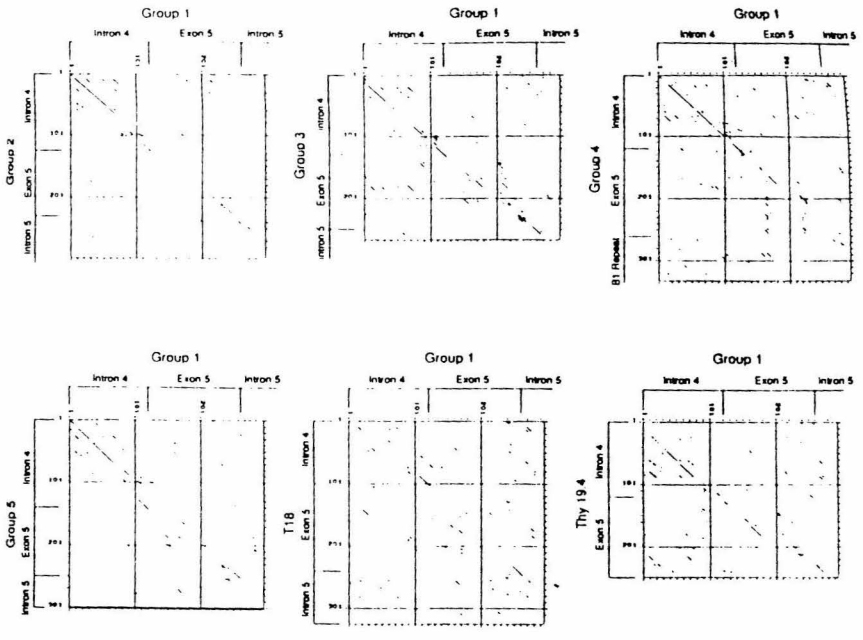


FIGURE 4. Dot matrix identity comparison of group 1 consensus sequences with groups 2–7 consensus sequences. Exon 5 boundaries are shown on the top and on the side. Each dot represents a six of eight nucleotide match between the sequences.

distinct breaks in similarity evident in the dot matrix identity plots suggest that the conservation in the fourth intron is a result of recombinational events that were involved in the evolution of the class I gene family. These recombination events could have included transmembrane exon shuffling or de novo generation events that created hybrid genes with similarity to classical class I genes in exon 4 and the 5′ portion of intron 4, but little or no similarity in exon 5 and the rest of intron 4. Alternatively, it is suggested that short introns in class I genes are generally more conserved than the interior portions of long introns because proposed recombination events that transfer exons between class I genes could often extend beyond the end of the exons into a portion of the surrounding introns (36). It is conceivable that the 5′ portions of the fourth introns are generally more conserved than the 3′ portions because such DNA segment exchange events could occur more often between fourth exons than fifth exons (14).

The exon 5 sequence data can be used to support models for the evolution of specific groups of class I genes. It is proposed that the $Q4^b$–$Q10^b$ genes in the C57BL/10 mouse resulted from duplications of a primordial *Qa* gene pair, with the even- and odd-numbered genes derived from one or the other of the primordial genes (37, 38). The *Q8* and *Q9* genes were subsequently fused in an unequal crossover event to form

the hybrid gene *Q8/9*^{d} of BALB/c mice (37). The exon 5 sequence data supports this model since exon 5 in the *Q4*^{d} and *Q6*^{d} genes share 97% similarity and a single nucleotide deletion causing a frame shift at nucleotide 62. In addition, the *Q7*^{d} and *Q8/9*^{d} genes are 99% similar to each other in exon 5. However, the *Q5*^{d} gene is 98% similar to the *D*^{d} gene in exon 5, suggesting that it had undergone a DNA segment exchange event from that gene.

The two gene clusters, *T1*^{c}*-T10*^{c} and *T11*^{c}*-37*^{c} (Fig. 1), of the *T1a* region of the BALB/c mouse, are proposed to have resulted from a duplication of an entire block of genes (14). The exon 5 sequence groups define gene pairs with representatives in the same order on both clusters. These gene pairs are *T1*^{c}*/T11*^{c}, *T2*^{c}*/T12*^{c}, *T3*^{c}*/T13*^{c}, *T6*^{c}*/T14*^{c}, *T7*^{c}*/T15*^{c}, *T8/T16*^{c}, *T9*^{c}*/T17*^{c}, and *T10*^{c}*/37*^{c}. The placement of these pairs in a specific order in both clusters supports the cluster duplication model. Because of similarities in restriction enzyme site patterns, the *T4*^{c} and *T5*^{c} genes are proposed to have been created in a duplication of a pair of genes that also produced the *T6*^{c} and *T7*^{c} genes (14). However, the exon 5 sequence data does not support this contention since the *T5*^{c} exon 5 does not resemble the *T7*^{c} exon 5, but instead is 95% similar to the *T2*^{c} exon 5 in the first 58 nucleotides. Beyond that point, it does not resemble any other exon 5 sequence. In addition, exon 5 in the *T4*^{c} gene is 93% similar to that of the *T6*^{c} gene in the first 46 nucleotides, after which is a 21-bp polythymidine tract followed by a nonhomologous sequence. Since both the *T4*^{c} and *T5*^{c} exon 5 sequences are interrupted by nonhomologous sequences, it is likely that instead of being created as a block duplication of the *T6*^{c} and *T7*^{c} genes, the *T4*^{c} and *T5*^{c} genes are partial class I genes that were duplicated separately from distinct sources. The *T5*^{c} gene is probably a partially duplicated *T2*^{c} or *T12*^{c} gene, while the *T4*^{c} gene is probably a partially duplicated group 1 class I gene.

Splice Junction Sequences. The putative acceptor splice junction sequences of 35 and donor sequences of 30 of the class I fifth exons are shown in Fig. 5. The fifth exons of the *T4*^{c}, *T5*^{c}, *T7*^{c}, and *T15*^{c} genes do not have donor splice sequences; probably because they were eliminated by recombination or repetitive element integration events during their evolution. In addition, since *Thy-19.4* transcripts do not splice exon 5 to any 3′ exons (11), it is also excluded from the donor sequence figure. All 35 acceptor sequences have polypyrimidine tracts of 16–58 bp in length, followed by an AG dinucleotide. Since splicing invariably occurs after an AG dinucleotide in eukaryotic genes (39), and polypyrimidine tracts in acceptor splice signals are generally >11 bp in length, all 35 acceptor splice sequences appear functional. Similarly, all of the donor splice sequences match the consensus sequence ($^{C}_{A}$AG/GT$^{A}_{G}$AGT) with at least six of nine nucleotides and have the invariant GT dinucleotide at the immediate splice junction. Since donor splice sequences only need to match the consensus sequence in as little as five of the nine nucleotides to be functional (40), and since all of the class I donor sequences have the invariant GT dinucleotide found in all eukaryotic donor sequences (39), all 30 of the class I donor sequences also appear to be functional. Since none of the splice sequences in the 35 class I fifth exons appear abnormal, it is unlikely that any of the fifth exons will be nonfunctional because of splicing abnormalities.

In addition to donor and acceptor splice sequences homologous to those in previously characterized genes, several possible alternative splice sites can be identified (Fig. 3). These sites include in-frame donor sequences in members of groups 1-5

FIGURE 5. The acceptor and donor splice sequences of BALB/c class I genes. The consensus sequence (Con) appears above the compiled sequences, and vertical lines indicate where splicing is predicted to occur.

in intron 4 near the junction with exon 4. If a class I transcript splices at these sites, between two and five amino acids would be added to the α3 domain of its translation product. In the *T2*^{c}, *T5*^{c}, *T12*^{c}, and *T18*^{c} genes, there are additional possible alternative donor sites in intron 4 that generate a different frame than that identified in cDNA transcripts. However, transcripts that splice these possible alternative donor sites to possible alternative acceptor sites in the 5′ portion of exon 5 would place the hydrophobic translation of exon 5 in the same reading frame as exon 4. The translation would be slightly longer in exon 4 and slightly shorter in exon 5. The translation of the *T12*^{c} fifth exon terminates two amino acids after the acceptor splice signal homologous to those characterized in other class I genes. However, a *T12*^{c} transcript could encode a hydrophobic transmembrane if spliced at the possible alternative splice sites. Finally, there are donor signals in the fifth exon of the *T4*^{c} gene. However, since these sites were probably introduced to this gene by a recombination event, it is unclear if they actually evolved to splice the *T4*^{c} fifth exon to a 3′ exon.

Analysis of Predicted Sequence. Translation of the DNA sequences reveals that, with the exception of the *Q10*^{d}, *T5*^{c}, *T11*^{c}, and *T12*^{c} genes, each class I gene has an open reading frame in exon 5, whose translation is potentially hydrophobic and long enough to span a lipid bilayer (27; Fig. 6). Thus, each of these 31 fifth exon–encoded amino acid sequences can be divided into connecting, transmembrane and cytoplasmic segments. In this study the transmembrane segment is arbitrarily defined as the most hydrophobic 21 amino acids of the fifth exon translation, since that is the chain length required to form an α helix that can completely span a lipid bilayer (41). Analysis of previously characterized membrane-bound proteins reveals that transmembrane domains range in length from 20 to 28 amino acids (42). Since the choice of 21 amino acids is arbitrary, it is important to note that it is possible that in the actual gene

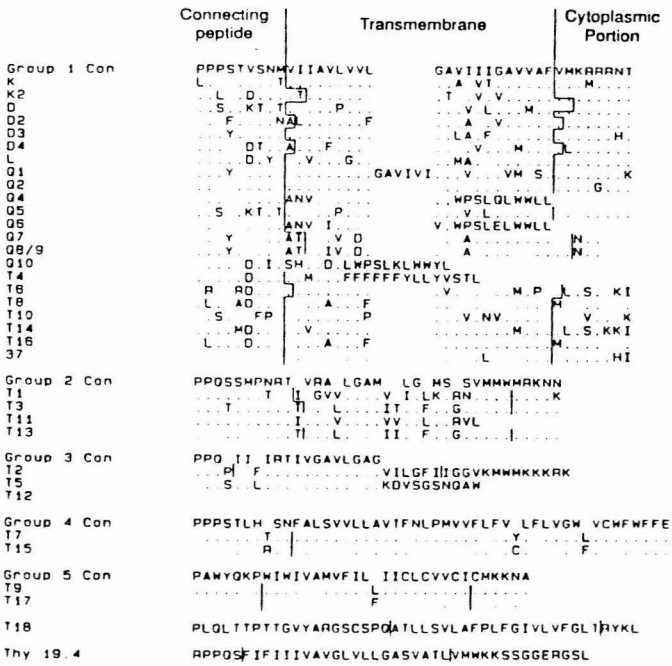


FIGURE 6 The translations of the exon 5 sequences. Where they exist, borders between predicted connecting, transmembrane, and cytoplasmic segments are indicated by vertical lines between amino acids. Frameshifts present in the fifth exons of the $Q4^d$, $Q6^d$, and $Q10^d$ genes cause a portion of their exon 5 translation to be nonhomologous to those of other class I genes in their group, while recombination events that occurred in the $T4^c$ and $T5^c$ fifth exons produced a similar result for their translations.

products some of the residues near the calculated borders may not reside in their predicted environment. The exceptions to the arbitrary assignment of 21 amino acids are the predicted $Q1^d$, $T7^c$, and $T15^c$ transmembrane domains, which clearly have hydrophobic segments in excess of 27 amino acids. In the translated sequences, the putative connecting peptides vary from 4 to 20 amino acids in length. The putative cytoplasmic portions are between 4 and 13 amino acids in length, except in the $Q4^d$, $Q6^d$, $T4^c$, and $T15^c$ transmembrane domains, where none can be identified.

Connecting Peptides. Analysis of the putative connecting peptides reveals that the amino acid usage is similar to that in the hinge regions of Igs (24; Table IV). Proline (28%) is the most commonly used amino acid. In addition, asparagine (8%) is also present in these segments. These amino acids tend to disrupt any helical structure that may form in the junction between the transmembrane and outer domains (43). In addition, serine and threonine predominate in the connecting peptide at 16 and 14%, respectively. These two amino acids with small polar hydroxyl side chains are common in exposed areas, and their presence is not predicted to contribute to or disrupt the formation of α helices (43). However, comparison of 31 proteins exhibiting

TABLE IV

Amino Acid Compositions of Predicted Connecting, Transmembrane, and Cytoplasmic Segments

Amino acid	Connecting segment	Transmembrane	Cytoplasmic portion
		%	
Acidic			
Aspartic acid (D)	2.5	0.3	0
Glutamic acid (E)	0	0.4	1.0
Basic			
Lysine (K)	1.4	0.1	19.0
Arginine (R)	2.8	0.6	25.1
Histidine (H)	0.7	0	1.0
Polar			
Glycine (G)	0.7	8.3	3.1
Asparagine (N)	8.2	0.9	10.8
Glutamine (Q)	3.2	0.1	0
Cysteine (C)	0.4	1.6	1.0
Serine (S)	16.0	1.7	3.1
Threonine (T)	14.2	1.6	6.7
Tyrosine (Y)	2.8	0.1	0.5
Nonpolar			
Alanine (A)	2.8	15.1	1.0
Valine (V)	4.6	26.5	7.7
Leucine (L)	2.8	12.9	5.6
Isoleucine (I)	0	14.2	2.1
Proline (P)	28.4	1.3	0
Phenylalanine (F)	0.7	7.8	0
Methionine (M)	7.1	3.5	11.3
Tryptophan (W)	0.7	2.8	1.0

The calculation reflects percent representation in a total of 31 connecting and transmembrane segments, and 27 cytoplasmic segments. The individual amino acids have been previously assigned to acidic, basic, polar, or nonpolar categories (58).

segment flexibility demonstrates that both serines and threonines tend to be concentrated in flexible segments (44). The serines and threonines in the connecting peptides could confer more flexibility to this segment.

The imposition of a flexible β-turn structure in the connecting peptide could facilitate stretching and pivoting at this segment in a manner similar to that in the hinge region of Igs. It is suggested that the freedom of movement of the two Ig Fab arms relative to the Fc stem results from proline-rich amino acid sequences within the hinge segment that favor flexibility (24, 45, 46). This freedom of movement of the Fab arms is believed to be important for the function of Igs (47). Similarly, in the case of transplantation antigens, freedom of movement in the connecting peptide could be important to facilitate interaction with the TCR. It is interesting that the T18$^{\zeta}$ molecule is predicted to have a connecting peptide 20 amino acids in length. It would be twice as long as those in the transplantation antigens, but it is unclear if there is any significance to this difference.

Transmembrane Segments. In the predicted transmembrane segments, four hydrophobic amino acids dominate: valine (26%), alanine (15%), isoleucine (14%), and leucine (13%) (Table IV). Phenylalanine is present at an intermediate level of 8%, but the other three hydrophobic amino acids, tryptophan, methionine, and proline, each constitute 4% or less of the transmembrane amino acids. Proline (1%) may be absent from the transmembrane segments because it may tend to disrupt the α-helical structure assumed by the hydrophobic amino acids in their aliphatic environment (48, 49), although it has been suggested, on the basis of work with bacterial transmembrane domain deletion mutants, that transmembrane domains are not always completely α helical (50). Tryptophan (3%) and methionine (4%) are used less often in proteins in general, and their lower usage in the transmembrane domains could reflect this (51). In addition, tryptophan is suggested to be more hydrophilic than previously believed, and is represented infrequently in other transmembrane segments (52). Glycine (8%) is the only nonhydrophobic amino acid found in abundance in the transmembrane segments. Glycine has a very small slightly polar side chain that would probably not significantly decrease the hydrophobicity of the transmembrane segments. Interestingly, the putative $Q1^d$, $T7^c$, and $T15^c$ molecules are predicted to have hydrophobic transmembrane segments of between 27 and 49 amino acids. Although these genes have not been shown to encode proteins, if they did, it would be interesting to see how such long hydrophobic segments are accommodated in the membrane.

Proteins with transmembrane domains that interact with other proteins within the lipid bilayer, as class II MHC molecules are proposed to (53), are suggested to do so because they contain short stretches within their membrane spanning segment that are sufficiently amphipathic to promote such interactions (27). To test whether any of the class I transmembrane segments, as well as four BALB/c class II transmembrane segments, A^d_α, A^d_β, E^d_α, and E^d_β, could interact within the membrane with other proteins, the hydrophobic moment, a measure of amphipathicity, was calculated for 11-amino acid stretches within them. The length of 11 amino acids corresponds to approximately three turns of an α helix, which is believed to be the typical amphipathic segment size that interacts noncovalently with other proteins. For each transmembrane segment, the 11-amino acid stretch with the highest hydrophobic moment was determined, and that hydrophobic moment value was plotted against the hydrophobicity value for the 11-amino acid stretch (Fig. 7). Whether an 11-amino acid stretch is predicted to be sufficiently amphipathic to promote interactions within the membrane depends on which empirically defined area within the graph its plotted point falls (27).

This analysis reveals that the plotted points of all four class II transmembranes fall within or near the area defined as multimeric transmembrane. Since class II molecules are dimeric on the cell surface, it is proposed that the transmembrane's amphipathicity and amino acid sequence conservation are consistent with the hypothesis that they are dimeric within the lipid bilayer as well (53). On the other hand, the algorithm predicts that several class I transmembrane domains are not sufficiently amphipathic to be predicted to interact with other proteins within the membrane. The transplantation antigens, K^d, D^d, and L^d, are heterodimers with β_2-microglobulin, a small polypeptide with no membrane-spanning segment. Therefore, the prediction that they do not have amphipathic transmembrane segments is consis-

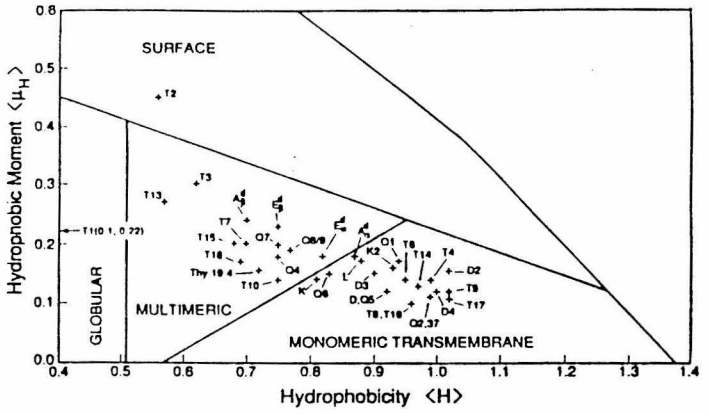


FIGURE 7. Hydrophobic moment plot. Hydrophobic moment (μ_H) is plotted against hydrophobicity (H) for 11-amino acid segments within each of 31 class I fifth exon translations, as well as the transmembrane domains of class II molecules A^d_α, A^d_β, E^d_α, and E^d_β. Each point plots within arbitrary areas labeled surface, globular, multimeric, or monomeric transmembrane. Although arbitrarily defined by Eisenberg et al. (27), 36 of 49 transmembrane segments originally used to define these regions were correctly plotted within the region of the graph that corresponded to their type.

tent with their probable monomericity within the membrane. In addition to the transplantation antigens, the putative molecules encoded by the majority of group 1 genes and the *T9*^{c} and *T17*^{c} genes are also predicted to be monomeric within the membrane. However, other putative class I transmembrane segments are predicted to be sufficiently amphipathic to associate within the membrane with other proteins. These include those predicted to be encoded by the *Q4*^{d}, *Q7*^{d}, *Q8/9*^{d}, *T3*^{c}, *T7*^{c}, *T10*^{c}, *T13*^{c}, *T15*^{c}, *T18*^{c}, and *Thy-19.4* genes. The *Q4*^{d} gene encodes a secreted class I product, and the *Q7*^{d} gene encodes the Qa-2 antigen, which is linked to the cell surface by a phosphatidylinositol linkage. The amphipathicity data is consistent with the hypothesis that during the processing or transport of these two molecules, intramembrane interactions occur with yet uncharacterized proteins. In addition, the putative products of several *T1a* region class I genes, including that of the *T13*^{c} gene, which encodes the TL^c antigen, are also predicted to interact with other proteins within the membrane. These *T1a* region genes, *T3*^{c}, *T7*^{c}, *T13*^{c}, *T15*^{c}, *T18*^{c}, and *Thy-19.4*, also have highly divergent fifth exons, suggesting that the molecules that these genes encode perform functions different than those of the classical class I molecules. The hydrophobic moment data is consistent with the hypothesis that their putative function may require intramembrane interactions with other molecules, possibly for the initiation of signaling cascades.

Cytoplasmic Segment. In the cytoplasmic portion, two basic amino acids predominate: arginine (25%) and lysine (19%). Basic amino acids are commonly found on the cytoplasmic side of transmembrane domains and are proposed to prevent the short cytoplasmic domain from being pulled through the hydrophobic lipid bilayer (42, 54). Histidine, a slightly basic amino acid, is not present in the class I anchor se-

quences. It is probably too weakly basic to serve in an anchor sequence. Also present in the cytoplasmic segments are methionine (11%), valine (8%), and asparagine (11%). It is unclear if there is any significance to the presence of these amino acids, although it is interesting that methionines and asparagines are clustered at both ends, but not at the center, of the transmembrane domains. Some of these residues may be spacers at the end of the domain and between the highly charged basic residues. Others are transmembrane segment amino acids included in cytoplasmic portion because of the arbitrary decision to limit the transmembrane segment to the 21 most hydrophobic amino acids.

Four of the 31 hydrophobic transmembrane domains do not have basic anchoring residues at the COOH-terminal end: $Q4^d$, $Q6^d$, $T4^c$, and $T15^c$. It is known that the *Q4*^{d} gene, like the *Q10*^{d} gene, encodes a secreted class I molecule (7, 55). The lack of an anchor sequence probably contributes to the fact that it is a secreted class I, in spite of its hydrophobic transmembrane segment. On the other hand, studies with *H-2L*^{d} gene mutants demonstrate that anchoring residues are not absolutely necessary for cell surface expression of class I glycoproteins (56). In addition, it is suggested that the $Q4^d$ molecule can exist on the cell surface of transfected cells (57). Clearly, the absence of anchoring residues can not universally be used as criteria for whether a class I molecule is secreted or membrane expressed. If there are products of the *Q6*^{d}, *T4*^{c}, and *T15*^{c} genes, it will be interesting to see whether they are secreted, membrane bound, or both. Interestingly, the putative transmembrane domains encoded by the *T7*^{c}, *T9*^{c}, and *T17*^{c} genes are predicted to end with only one or two basic anchoring residues, whereas the classical transplantation antigens are anchored by three to four basic residues (Fig. 6). If these genes can encode class I molecules, it will also be interesting to see if these molecules are anchored to the cell membrane as efficiently as the transplantation antigens. The $Q10^d$ molecule has neither a hydrophobic transmembrane nor anchoring residues (55), and it is known to be secreted (6).

Implication of Predicted Protein Sequences. This analysis reveals that almost all of the 35 class I genes have fifth exons that have open reading frames that could potentially encode a domain that is sufficiently hydrophobic and long to span a lipid bilayer, and hence by this criterion, appear to be functional. Only four of the 35 BALB/c class I gene fifth exons, those of the *Q10*^{d}, *T5*^{c}, *T11*^{c}, and *T12*^{c} genes, appear to be exceptions and have stop codons or frame shifts that prevent them from encoding a hydrophobic transmembrane domain. However, this does not necessarily imply that these four genes are pseudogenes, since at least the *Q10*^{d} gene encodes a presumably functional soluble class I molecule. Thus, based on the analysis of these sequences, it is not evident that any of these genes are pseudogenes. Of the remaining 31 class I genes, 27 have a fifth exon that could encode a domain similar to those of transplantation antigens in that it has both hinge-like connecting peptides and basic anchor amino acids at the appropriate ends of the hydrophobic stretch. Only the $Q4^d$, $Q6^d$, $T4^c$, and $T15^c$ transmembrane domains are exceptions by lacking basic anchor amino acids. Overall, the amino acid usage of these segments is appropriate for their predicted function. The hinge-like segments use amino acids expected to introduce β turns and segmental flexibility. The transmembrane segments consist of hydrophobic amino acids, while there are anchoring basic residues in the cytoplasmic segments. The maintenance of this motif is particularly striking because

of the extensive sequence divergence of the fifth exon groups. Since the majority of the fifth exons appear to be able to encode a functional transmembrane domain, it is unlikely that their divergence is merely a result of genetic drift in the absence of selective pressure. It is more likely that selective pressure exists to maintain them, suggesting that the majority of the class I genes, including the divergent ones, are functionally important. It could be speculated that the fifth exons have diverged from each other because the molecules that they encode have specialized functions other than antigen presentation to T cells. If the molecules that the divergent groups encode are involved in other functions, the fifth exons would still be selected for the ability to encode a transmembrane domain, but not one similar to those in restriction elements. Clearly, analysis of exon 5 sequences alone can only suggest what a particular class I gene can encode; further sequence and expression studies will be required to determine the extent of expression of class I genes.

Summary

DNA sequences of the fifth exon, which encodes the transmembrane domain, were determined for the BALB/c mouse class I MHC genes and used to study the relationships between them. Based on nucleotide sequence similarity, the exon 5 sequences can be divided into seven groups. Although most members within each group are at least 80% similar to each other, comparison between groups reveals that the groups share little similarity. However, in spite of the extensive variation of the fifth exon sequences, analysis of their predicted amino acid translations reveals that only four class I gene fifth exons have frameshifts or stop codons that terminate their translation and prevent them from encoding a domain that is both hydrophobic and long enough to span a lipid bilayer. Exactly 27 of the remaining fifth exons could encode a domain that is similar to those of the transplantation antigens in that it consists of a proline-rich connecting peptide, a transmembrane segment, and a cytoplasmic portion with membrane-anchoring basic residues. The conservation of this motif in the majority of the fifth exon translations in spite of extensive variation suggests that selective pressure exists for these exons to maintain their ability to encode a functional transmembrane domain, raising the possibility that many of the nonclassical class I genes encode functionally important products.

We thank Drs. I. Stroynowski, M. Zuniga, K. Fischer Lindahl, and J. Kobori for critically reviewing this manuscript; Dr. J. Howard for critical insights on exon 5 evolution and sharing unpublished rat exon 5 sequence data; and Mrs. C. Blagg for expert secretarial assistance.

Received for publication 19 June 1989 and in revised form 22 August 1989.

References

1. Silver, J., and L. Hood. 1974. Detergent solubilized H-2 alloantigen is associated with a small molecular weight polypeptide. *Nature (Lond.).* 249:764.
2. Zinkernagel, R. M., and P. C. Doherty. 1980. MHC-restricted cytotoxic T cells: studies on the role of polymorphic major transplantation antigens determining T-cell restriction specificity, function, and responsiveness. *Adv. Immunol.* 27:51.
3. Old, L. J., E. A. Boyse, and E. Stockert, E. 1963. Antigenic properties of experimental leukemias. I. Serological studies *in vitro* with spontaneous and radiation-induced leukemias. *J. Natl. Cancer Inst.* 31:977.

4. Flaherty, L. 1976. The *Tla* region of the mouse: identification of a new serologically defined locus, *Qa-2*. *Immunogenetics*. 3:533.
5. Stanton, T. H., and E. A. Boyse. 1976. A new serologically defined locus, *Qa-1*, in the *Tla* region of the mouse. *Immunogenetics*. 3:525.
6. Maloy, W., J. Coligan, Y. Barra, and G. Jay. 1984. Detection of a secreted form of the murine H-2 class I antigen with an antibody against its predicted carboxyl terminus. *Proc. Natl. Acad. Sci. USA*. 81:1216.
7. Robinson, P. J. 1985. Qb-1, a new class I polypeptide encoded by the *Qa* region of the mouse H-2 complex. *Immunogenetics*. 22:285.
8. Steinmetz, M., A. Winoto, K. Minard, and L. Hood. 1982. Clusters of genes encoding mouse transplantation antigens. *Cell*. 28:489.
9. Winoto, A., M. Steinmetz, and L. Hood. 1983. Genetic mapping in the major histocompatibility complex by restriction enzyme polymorphism: most mouse class I genes map to the *Tla* complex. *Proc. Natl. Acad. Sci. USA*. 80:3425.
10. Richards, C. S., M. Bucan, K. Brorson, M. Kiefer, S. Hunt, H. Lehrach, and K. Fischer Lindahl. Genetic and molecular mapping of the *Hmt* region of the mouse. *EMBO (Eur. Mol. Biol. Organ.) J.* In press.
11. Brorson, K., S. Richards, S. Hunt, H. Cheroutre, K. Fischer Lindahl, and L. Hood. 1989. Analysis of a new class I gene mapping to the *Hmt* region of the mouse. *Immunogenetics*. 30:273.
12. Goodenow, R., M. McMillan, M. Nicolson, B. Sher, K. Eakle, N. Davidson, and L. Hood. 1982. Identification of the class I genes of the mouse major histocompatibility complex by DNA-mediated gene transfer. *Nature (Lond.)*. 300:231.
13. Stephan, D., H. Sun, K. Fischer Lindahl, E. Meyer, G. Hämmerling, L. Hood, and M. Steinmetz. 1986. Organization and evolution of *D* region class I genes in the mouse major histocompatibility complex. *J. Exp. Med*. 163:1227.
14. Fisher, D. A., S. W. Hunt, and L. Hood. 1985. Structure of a gene encoding a murine thymus leukemia antigen, and organization of *Tla* genes in the BALB/c mouse. *J. Exp. Med*. 162:528.
15. Stroynowski, I., M. Soloski, M. Low, and L. Hood. 1987. A single gene encodes soluble and membrane-bound forms of the major histocompatibility Qa-2 antigen: anchoring of the product by a phospholipid tail. *Cell* 50:759.
16. Robinson, P., D. Bever, A. Mellor, and E. Weiss. 1988. Sequence of the mouse *Q4* class I gene and characterization of the gene product. *Immunogenetics*. 27:79.
17. Steinmetz, M., K. W. Moore, J. Frelinger, B. Sher, F. Shen, E. A. Boyse, and L. Hood. 1981. A pseudogene homologous to mouse transplantation antigens: transplantation antigens are encoded by eight exons that correlate with protein domains. *Cell*. 25:683.
18. Kvist, S., L. Roberts, and B. Dobberstein. 1983. Mouse histocompatibility genes: structure and organization of a K^d gene. *EMBO (Eur. Mol. Biol. Organ.) J.* 2:245.
19. Blobel, G., and B. Dobberstein. 1975. Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane bound ribosomes of murine myeloma. *J. Cell Biol*. 67:835.
20. Uehara, H., J. Coligan, and S. Nathenson. 1981. Isolation and sequence analysis of the intramembranous hydrophobic segment of the H-$2K^b$ murine histocompatibility antigen. *Biochemistry*. 20:5936.
21. Bjorkman, P., M. Saper, B. Samraoui, W. Bennett, J. Strominger, and D. Wiley. 1987. Structure of the human class I histocompatibility antigen, HLA-A2. *Nature (Lond.)*. 329:506.
22. Sanger, F., A. R. Coulson, B. G. Barrell, A. J. H. Smith, and B. A. Roe. 1980. Cloning in single stranded bacteriophage as an aid to rapid DNA sequencing. *J. Mol. Biol*. 143:161.
23. Moore, K. W., B. Sher, Y. H. Sun, K. A. Eakle, and L. Hood. 1982. DNA sequence

of a gene encoding a BALB/c mouse L^d transplantation antigen. *Science (Wash. DC).* 215:679.

24. Kabat, E., T. Wu, M. Reid-Miller, H. Perry, and K. Gottesman. 1977. Sequences of Proteins of Immunological Interest, 4th ed. U.S. Department of Health and Human Services, Public Health Services, National Institutes of Health, Bethesda, MD.
25. Needleman, S., and C. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48:443.
26. Kimura, M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA.* 78:454.
27. Eisenberg, D., E. Schwarz, M. Komaromy, and R. Wall. 1984. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* 179:125.
28. Eisenberg, D., R. Weiss, T. Terwilliger, and W. Wilcox. 1982. Hydrophobic moments and protein structure. *Faraday Symp. Chem. Soc.* 17:109.
29. Steinmetz, M., J. Frelinger, D. Fisher, T. Hunkapiller, D. Pereira, S. Weissman, H. Uehara, S. Nathenson, and L. Hood. 1981. Three cDNA clones encoding mouse transplantation antigens: homology to immunoglobulin genes. *Cell.* 24:125.
30. Chen, Y.-T., Y. Obata, E. Stockert, and L. Old. 1985. Thymus-Leukemia (TL) antigens of the mouse. Analysis of TL mRNA and TL cDNA from TL^+ and TL^- strains. *J. Exp. Med.* 162:1134.
31. Krayev, A., D. Kramerov, K. Skryabin, A. Ryskov, A. Bayev, and G. Georgiev. 1980. The nucleotide sequence of the ubiquitous repetitive DNA sequence B1 complementary to the most abundant class of mouse fold-back RNA. *Nucleic Acids Res.* 8:1201.
32. Crews, S., J. Griffin, J. Huang, K. Calame, and L. Hood. 1981. A single V_H gene segment encodes the immune response to phosphoryl choline: Somatic mutation is correlated with the class of antibody. *Cell.* 25:59.
33. Kastern, W. 1985. Characterization of two class I major histocompatibility rat cDNA clones, one of which contains a premature termination codon. *Gene (Amst.).* 34:227.
34. Davis, N., and P. Model. 1985. An artificial anchor domain: hydrophobicity suffices to stop transfer. *Cell.* 41:607.
35. Klein, J., and F. Figueroa. 1986. The evolution of class I MHC genes. *Immunol. Today.* 7:41.
36. Hayashida, H., and T. Miyata. 1983. Unusual evolutionary conservation and frequent DNA segment exchange in class I genes of the major histocompatibility complex. *Proc. Natl. Acad. Sci. USA.* 80:2671.
37. Weiss, E. H., L. Golden, K. Fahrner, A. L. Mellor, J. J. Devlin, H. Bullman, H. Tiddens, H. Bud, and R. A. Flavell. 1984. Organization and evolution of the class I gene family in the major histocompatibility complex of the C57BL/10 mouse. *Nature (Lond.).* 310:650.
38. Devlin, J., E. Weiss, M. Paulson, and R. Flavell. 1985. Duplicated gene pairs and alleles of class I genes in the *Qa* region of the murine major histocompatibility complex: a comparison. *EMBO (Eur. Mol. Biol. Organ.) J.* 4:3203.
39. Breathnach, R., and P. Chambon. 1981. Organization and expression of eukaryotic split genes coding for proteins. *Annu. Rev. Biochem.* 50:349.
40. Mount, S. 1982. A catalogue of splice junction sequences. *Nucleic Acids Res.* 10:459.
41. Tanford, C. 1980. The hydrophobic effect: Formation of Micelles and Biological Membranes. Wiley, New York.
42. Warren, G. 1981. Membrane proteins: structure and assembly. *In* Membrane Structure. J. B. Finean, and R. H. Michell, editors. Elsevier Science Publishers B. V., Amsterdam. 215–257.
43. Chou, P., and G. Fasman. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.* 47:45.
44. Karplus, P., and G. Schulz. 1985. Prediction of chain flexibility in proteins. *Naturwissen-*

schaften. 72:212.

45. Seegan, G., C. Smith, and V. Schumaker. 1979. Changes in quaternary structure of IgG upon reduction of the interheavy-chain disulfide bond. *Proc. Natl. Acad. Sci. USA.* 76:907.
46. Marquart, M., J. Deisenhofer, R. Huber, and W. Palm. 1980. Crystallographic refinement and atomic models of the intact immunoglobulin molecule Kol and its antigen-binding fragment at 3.0Å and 1.9Å resolution. *J. Mol. Biol.* 141:369.
47. Klein, M., N. Haeffner-Cavaillon, D. Isenman, C. Rivat, M. Navia, P. Davies, and K. Dorrington. 1981. Expression of biological effector functions by immunoglobulin G molecules lacking the hinge region. *Proc. Natl. Acad. Sci. USA.* 78:524.
48. Henderson, R., and P. N. T. Unwin. 1975. Three dimensional model of purple membrane obtained by electron microscopy. *Nature (Lond.).* 257:28.
49. Guidotti, G. 1977. The structure of intrinsic membrane proteins. *J. Supramol. Struct.* 7:489.
50. Davis, N., J. Boeke, and P. Model. 1985. Fine structure of a membrane anchor domain. *J. Mol. Biol.* 181:111.
51. Klapper, M. 1977. The independent distribution of amino acid near neighbor pairs into polypeptides. *Biochem. Biophys. Res. Commun.* 78:1018.
52. Clothia, C. 1976. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* 105:1.
53. Malissen, M., T. Hunkapiller, and L. Hood. 1983. Nucleotide sequence of a light chain gene of the mouse *I-A* subregion: A^d_β. *Science (Wash. DC).* 221:750.
54. Tomita, M., and V. Marchesi. 1975. Amino acid sequence and oligosaccharide attachment sites of human erythrocyte glycophorin. *Proc. Natl. Acad. Sci. USA.* 72:2964.
55. Cosman, D., M. Kress, G. Khoury, and G. Jay. 1982. Tissue specific expression of an unusual *H-2* (class I)-related gene. *Proc. Natl. Acad. Sci. USA.* 79:4947.
56. Zuniga, M., and L. Hood. 1986. Clonal variation in cell surface display of an H-2 protein lacking a cytoplasmic tail. *J. Cell Biol.* 102:1.
57. Schepart, B., J. Woodward, M. Palmer, M. Macchi, P. Basta, E. McLaughlin-Taylor, and J. Frelinger. 1985. Expression in L cells of transfected class I genes from the mouse major histocompatibility complex. *Proc. Natl. Acad. Sci. USA.* 82:5505.
58. Alberts, B., D. Bray, J. Lewis, M. Rall, K. Roberts, and J. Watson. 1983. Molecular Biology of the Cell. Garland Publishing, New York.
59. Sher, B., R. Nairn, J. E. Coligan, and L. Hood. 1985. DNA sequence of the mouse *H-2D*^{d} transplantation antigen gene. *Proc. Natl. Acad. Sci. USA.* 82:1175.
60. Transy, C., S. R. Nash, B. David-Watine, M. Cochet, S. W. Hunt, L. E. Hood, and P. Kourilsky. 1987. A low polymorphic mouse *H-2* class I gene from the *Tla* complex is expressed in a broad variety of cell types. *J. Exp. Med.* 166:341.
61. Fisher, D. A., M. Pecht, and L. Hood. 1989. DNA sequence of a class I pseudogene from the *Tla* region of the murine MHC: recombination at a B2 Alu repetitive sequence. *J. Mol. Evol.* 28:306.
62. Headly, M., S. Hunt, K. Brorson, J. Andris, L. Hood, J. Forman, and P. Tucker. 1989. DNA sequence analysis of *D2*^{d}: a new *D*-region class I gene. *Immunogenetics.* 29:359.

VI.4 Immunoglobulins

The joining of V and J gene segments creates antibody diversity

Martin Weigert, Robert Perry & Dawn Kelley

The Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, Pennsylvania 19111

Tim Hunkapiller, James Schilling & Leroy Hood

California Institute of Technology, Division of Biology, Pasadena, California 91125

The variable regions of mouse kappa (κ) chains are coded for by multiple variable (V) gene segments and multiple joining (J) gene segments. The V_κ gene segments code for residues 1 to 95; the J_κ gene segments code for residues 96 to 108 (refs 1–3). This gene organisation is similar to that encoding the V_λ regions[4]. Diversity in V_κ regions arises from several sources: (1) there are multiple germ-line V_κ gene segments and J_κ gene segments; (2) combinatorial joining of V_κ gene segments with different germ-line J_κ gene segments; and possibly, (3) somatic point mutation, as postulated for V_λ gene segments[5]. Also, from a comparison of the number of germ-line J_κ gene segments and amino acid sequences, it has been suggested that J_κ region sequences may be determined by the way V_κ and J_κ gene segments are joined[2,3]. This report supports this model by directly associating various J_κ sequences with given J_κ gene segments.

The number of germ-line J_κ gene segments has been estimated in two ways. A region of DNA containing J_κ gene segments has been characterised by DNA sequence analysis and was found to include five J_κ (or J_κ-like) gene segments, only four of which have been found to be expressed as J segments[2,3]. Additional J_κ gene segments have not been found, although their presence cannot be excluded rigorously. Alternatively, the number of J_κ gene segments may be determined by counting the number of distinctive RNA precursor patterns associated with different κ chain mRNAs[6]. Since the κ-mRNA precursors contain the intervening sequence between the J_κ and C_κ segments, they have a characteristic size depending on which J segment is used, the largest precursor corresponding to the J_κ segment most distal to the C_κ gene and the smallest to the most proximal J_κ segment. On this basis the RNA precursor patterns for 20 different myeloma tumours were classified into four different categories, corresponding to the four different J_κ gene segments that these chains use. The distances between the C_κ gene and the various J_κ gene segments deduced from the analysis of precursor patterns in NZB plasmacytomas agree with the distances determined by restriction mapping and DNA sequence analysis of BALB/c germ-line DNA. A previously suspected discrepancy[6] has been resolved by a more accurate calibration of the mobility data.

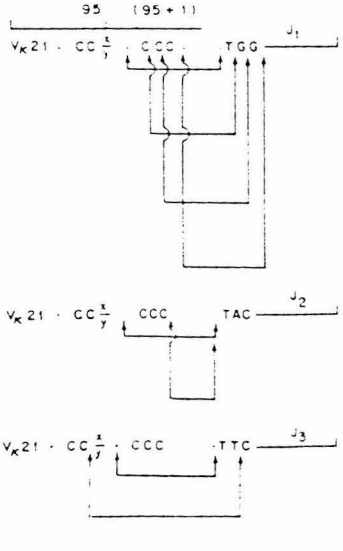
We have analysed κ chains of the V_κ 21 group from the NZB mouse strain. To date ten different J_κ amino acid sequences have been observed in twenty V_κ 21 examples (Fig. 1). Thus the

	Sequences of J segments associated with V_κ21 segments		Frequency	Processing category
Germline	96	107		
J_1	W T F G G G T K L E I K R		5/20	I
J_2	Y———		3/20	II
J_3	F———S———		2/20	III
J_4	L———A———L———		4/20	IV
$J_?$	W———D———E———		1/20	I
Variant				
J_1** PC6684	R———		1/20	I
J_1** PC7940	P———		1/20	I
J_2** PC7132	(P)Y———		1/20	II
J_3** PC8701	[ ]———S———		1/20	III
J_4** PC8982	L———A———E———L———		1/20	IV

Fig. 1 J segments associated with V_κ21 chains from NZB myeloma antibodies. J segments are numbered according to their processing category[5]. This is based on the relative sizes of the κ-mRNA precursors, as determined by gel-electrophoretic analysis of poly(A)$^+$ nuclear RNA. These precursors are approximately 5.3, 5.0, 4.4 and 4.1 kilobases for categories I, II, III and IV, respectively. The amino-terminal sequences (40 residues) of PC8701, PC8982 and PC10916 were determined on a modified Beckman sequenator as described previously[11] and were found to be representatives of the V_κ21 subgroups, D, C and C, respectively. The J segment sequence of these κ chains were determined as follows: each κ chain was subjected to mild acid cleavage[12] of the aspartylproline between residues 94 and 95. The peptide mixture was sequenced without separation after the amino-terminal peptide was blocked by addition of fluorescamine[13]. The rest of the sequences are taken from ref. 1. The germ-line J segments are defined as described in the text. J_1, J_2, J_3 and J_4 are coded for by the BALB/c embryo J gene segments termed J_1, J_2, J_4 and J_5 in ref. 2 or those termed J_5, J_4, J_3 and J_1 in ref. 3. The PC2413 J segment may be coded for by a germ-line J gene or variant of a J gene unique to the NZB mouse.

number of J_κ sequences appears to be greater than the number of germ-line J_κ gene segments. There are three types of J_κ sequence. First, four distinct types (J_1, J_2, J_3 and J_4) correspond to those coded for by four of the germ-line J_κ gene segments in BALB/c DNA. Each of these J_κ sequences has been observed two or more times in the NZB V_κ21 regions (Fig. 1). These J_κ sequences also have been observed in BALB/c κ chains[7] indicating that NZB and BALB/c germ-line J_κ gene segments code for the same J_κ regions. A second type of J_κ sequence differs by a single residue from those of the germ-line J_κ gene segments; these are termed variant J_κ sequences. There are five J_κ sequences of this type and each has been observed only once.

There are two possible explanations for the variant J_κ sequences: (1) they arise from somatic diversification which occurs at the junctional region during the joining of V_κ and J_κ gene segments or (2) they are encoded by additional germ-line J_κ gene segments that are absent from or have not yet been detected in the BALB/c genome. The former model predicts that the precursors of the mRNAs coding for these variant J_κ segments would belong to a category typical of one of the four germ-line examples; the latter model predicts a unique mRNA precursor for each variant. A third type of J sequence which occurs in the exceptional PC2413, differs from the other germ-line J_κ gene segments by at least two nucleotide substitutions.



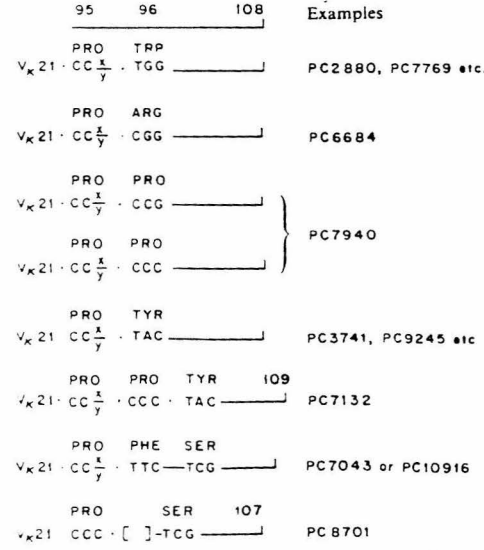


Fig. 2 Fusion of V_κ21 genes and J gene segments. The fusion can occur to join V and J gene segments exactly to form a contiguous V–J gene. Such genes will have germ-line J region sequences as in the case of PC2880, PC3741 or PC7043 and other examples[1]. Fusion to J genes might include nucleotides in the intervening sequence (95 + 1). If this codon is CCC for all V_κ21, then the variant sequences, PC6684, PC7940 and PC7132 can be explained by fusion to a J gene (PC7132), within a J gene (PC6684 and PC7940) or within a V gene (PC8701).

Differences of this extent raise the issue of whether the PC2413 J_κ sequence might be encoded by a fifth germ-line J_κ gene segment in the NZB mouse.

An analysis of the RNA processing patterns for the tumours producing κ chains with variant J_κ sequences indicates that each pattern belongs to one of the four categories characteristic of germ-line J_κ gene segments (Fig. 1). Thus it appears that these variant J sequences arise during the process of V_κ–J_κ joining.

Knowing the size of the mRNA precursor for a J_κ variant allows us to infer the kind of somatic mutation that could have produced the variant sequence. For example, both PC6684 and PC7940 mRNAs exhibit the J_1 type prescursor. Their J segments differ from the germ-line J_1 segment at position 96, presumably due to modifications of the tryptophan codon 96. PC7132 arose from the insertion of a proline codon before the J_2 gene segment. PC8701 resulted from a deletion of the phenylalanine codon at the beginning of J_3 gene segment. In addition to this junctional diversity, random point mutations either in the germ line or during somatic differentiation may occasionally occur (for example, PC8982).

The κ-mRNA precursor in PC2413 appears to belong to the category characteristic of the germ-line J_1 segment. Thus, if the J sequence of this tumour is actually encoded by a unique NZB germ-line J_κ segment, thus segment would have to be located very close to the J_1 segment, or alternatively, be associated with a second, as yet unidentified, C_κ gene.

The nature of most J_κ variants suggests that they may be due to variations arising from the V–J joining process because the substitutions, mutations, insertions or deletions generally occur at the beginning of J_κ segments. These variations may be explained by two base substitutions (for example, tryptophan to proline) as well as the insertion or deletion of codons. Thus this variation does not appear to arise merely by single-base somatic mutation. Leder and Tonegawa and their coworkers have recently observed identical inverted repeat sequences at the 3′ end of the V gene segment and at the 5′ end of the J gene segment. They have hypothesised that these inverted repeats may form paired stem structures that juxtapose the V and J gene segments and that recombination may occur across the base of this stem at slightly different points to generate V–J junctional diversity (refs 2 and 3 and Fig. 2). This site-specific recombinational model appears to explain the V–J junctional diversity that we have noted in the $V_\kappa 21$ chains.

By whatever mechanism V gene segments are translocated to J gene segments, a continuous DNA sequence coding for the entire V region must be generated[4]. Since all mouse κ chains studied to date have proline at residue 95 (CC^{a}_{y}), a V_κ gene segment translocation to the J_1 gene segment (tryptophan (TGG) at residue 96) will result in the contiguous DNA sequence, CC^{a}_{y} TGG at the V–J joining site (Fig. 2). Such a translocation will create a V_κ gene segment that is associated with the J_1 germ-line gene segment. This type of translocation explains five out of the twenty $V_\kappa 21$ regions observed (Fig. 1). Likewise, similar translocations between $V_\kappa 21$ gene segments and the J_2, J_3 and J_4 gene segments can explain the nine additional examples of germ-line J segments (Fig. 1).

The variant J segments may arise by recombination within rather than between the V gene segment codons for positions 95 or 95 + 1 and their counterparts in the J gene segments (Fig. 2). Codon 95 + 1 is CCC in several V_κ gene segments[2,8]. If the (95 − 1) codon is CCC for all $V_\kappa 21$ gene segments, the type of variation that may occur at the V–J joining site is limited. Amino acid substitutions at residue 96 for a $V_\kappa 21$–J_1 translocation can only be arginine (CGG) or proline (CCG) (Fig. 2). An amino acid insertion would result in an extra proline codon before the J gene segment. In addition, deletions of residue 96 could occur. As shown in Fig. 2, the variants observed so far fit the predictions of this model. If there is variation at the (95 + 1) codon, a greater variety of substitutions could occur at residue 96[3]. Obviously similar variation could occur in the (96 − 1) codon of the J gene segment. Additional analyses of $V_\kappa 21$ chains are under way to define more completely the amino acid sequence variability at the V–J junctional site.

In our survey of $V_\kappa 21$ chains, the frequency of κ chains expressing variant J segments relative to those expressing germline J segments is high (5 out of 20). This frequency could be high because the N-terminus of the J_κ segment is important for combining-site size and shape[9] and antigen selection of specific antibodies amplifies mutations that occur at a low frequency. (The uniformity of J_λ segments[10] may mean that variation in this region does not affect the complementarity of antibodies with λ chains.) Alternatively, this frequency may be a true reflection of the rate of variation by this mechanism.

This research was supported by a grant from the NSF, grants from the NIH and an appropriation from the Commonwealth of Pennsylvania.

Received 4 September; accepted 11 November 1979.

1. Weigert, M., Gatmaitan, L., Loh, E., Schilling, J. & Hood, L. *Nature* **276**, 785–790 (1978).
2. Sakano, H., Huppi, K., Heinrich, G. & Tonegawa, S. *Nature* **280**, 288–294 (1979).
3. Max, E. E., Seidman, J. G. & Leder, P. *Proc. natn. Acad. Sci. U.S.A.* **76**, 3450–3454 (1979).
4. Brack, C., Hirama, M., Lenhard-Schuller & Tonegawa, S. *Cell* **15**, 1–14 (1978).
5. Weigert, M., Cesari, I. M., Yonkovich, S. J. & Cohn, M. *Nature* **228**, 1045–1047 (1970).
6. Perry, R. P., Kelley, D. E. & Schibler, U. *Proc. natn. Acad. Sci. U.S.A.* **76**, 3678–3682 (1979).
7. Kabat, E. A., Wu, T. T. & Bilofsky, H. (eds) *Variable Regions of Immunoglobulin Chains. Tabulations and Analysis of Amino Acid Sequences* (Bolt, Baranek & Newman, Massachusetts, 1976).
8. Seidman, J. G., Leder, A., Nau, M., Norman, B. & Leder, P. *Science* **202**, 11–17 (1978).
9. Padlan, E. A., Davies, D. R., Rudikoff, S. & Potter, M. *Immunochemistry* **13**, 945–949 (1976).
10. Weigert, M. & Riblet, R. *Cold Spring Harb. Symp. quant. Biol.* **41**, 837–846 (1976).
11. Hunkapiller, M. & Hood, L. *Biochemistry* **17**, 2124–2133 (1978).
12. Paulsen, K., Fraser, K. J. & Haber, E. *Proc. natn. Acad. Sci. U.S.A.* **69**, 2495–2499 (1972).
13. Udenfriend, S. *et al. Science* **178**, 871–872 (1972).

0028-0836/80/050499-03$01.00

Proc. Natl. Acad. Sci. USA
Vol. 78, No. 3, pp. 1800–1804, March 1981
Immunology

Expression of IgD may use both DNA rearrangement and RNA splicing mechanisms

(IgD myeloma tumor/δ gene organization/δ gene expression)

K. W. MOORE*, J. ROGERS†, T. HUNKAPILLER*, P. EARLY*, C. NOTTENBURG‡, I. WEISSMAN‡, H. BAZIN§, R. WALL†, AND L. E. HOOD*

*Division of Biology, California Institute of Technology, Pasadena, California 91125, † Molecular Biology Institute and Department of Microbiology and Immunology, UCLA School of Medicine, Los Angeles, California 90024, ‡ Laboratory of Experimental Oncology, Department of Pathology, Stanford Medical School, Stanford, California 94305, and § Experimental Immunology Unit, University of Louvain, Brussels, Belgium.

Communicated by Norman Davidson, November 4, 1980

ABSTRACT From a library of mouse sperm DNA, we have isolated two overlapping clones which contain the C_δ gene. One of these clones also contains the C_μ gene. The C_δ gene is separated from the C_μ membrane exons by approximately 2 kilobases (kb) of DNA. The C_δ gene was identified by (*a*) hybridization to poly(A)$^+$RNA prepared from the IgD-producing rat plasma cell tumor IR731, and (*b*) homology of a translated nucleotide sequence to the amino acid sequence of the human δ chain. The C_δ gene spans 8 kb of DNA in the germ line. Plasmid subclones of the C_δ gene were used as probes in Southern and RNA blot experiments. RNA blot analysis of cytoplasmic poly(A)$^+$RNA from IR731 and a $\mu^+\delta^+$ B-cell hybridoma revealed 1.6- and 2.7-kb δ mRNA species with different 3′ ends, which presumably encode the secreted and membrane-bound forms, respectively, of the δ chain. Southern blot analysis of DNA from two $\mu^+\delta^+$ lymphomas revealed that the C_δ gene is in the germ-line configuration in each case. Restriction map analysis of C_μ and C_δ genomic clones isolated from a library of normal $\mu^+\delta^+$ B-cell DNA also gave no evidence for DNA rearrangement in the region between the C_μ and C_δ genes. Taken together, these data suggest that IgD expression in $\mu^+\delta^+$ B cells does not involve a V_H-to-C_δ DNA switch rearrangement. We propose that simultaneous expression of C_μ and C_δ with a single V_H gene is mediated by two alternative routes of RNA processing of a primary nuclear transcript which contains the V_H, C_μ, and C_δ genes. In contrast, analogous experiments with myeloma IR731 DNA revealed that the C_μ gene has been deleted from the myeloma DNA and that the C_δ gene has undergone DNA rearrangement, presumably including a switch recombination of the V_H gene from the C_μ to the C_δ gene. These results indicate that two alternative mechanisms may be used in the expression of IgD molecules—RNA splicing in B cells and DNA rearrangement in plasma cells.

An immunoglobulin heavy chain is composed of a variable (V_H) region and one of five classes of constant (C_H) region: C_μ (IgM), C_δ (IgD), C_γ (IgG), C_α (IgA), and C_ϵ (IgE). In mice, the five C_H classes are encoded by eight distinct genes: C_μ, C_δ, $C_{\gamma 1}$, $C_{\gamma 2a}$, $C_{\gamma 2b}$, $C_{\gamma 3}$, C_α, and C_ϵ. Early in its development, a lymphocyte (B cell) bears only IgM on its surface; later, IgD molecules are often expressed together with IgM molecules (for review, see ref. 1). Upon interaction with antigen, a B cell proliferates and differentiates, ultimately becoming a plasma cell. The class of immunoglobulin produced by its progeny may change, from IgM (and IgD) to IgG, IgA, or IgE.

The molecular mechanisms by which these genes are expressed during B-cell development have been partially characterized. The V_H gene is joined to the C_μ gene by assembly of three gene segments: *V*, *D*, and *J* (2, 3). Presumably, upon recognition of antigen, the V_H gene, along with some C_μ 5′-flanking sequence, is joined to another C region in a phenomenon called the heavy-chain switch (4–6). Honjo and coworkers (7, 8) have suggested that the intervening DNA, containing the C_μ gene and perhaps other C_H genes, is deleted.

Both IgM and IgD molecules are present on the surface of the B cell. Experiments utilizing allotypic markers have shown that their expression conforms to the rule of allelic exclusion: both heavy chains on the surface of an individual cell are encoded by the same chromosome (9). Moreover, considerable evidence suggests that these two cell-surface molecules bear identical V_H regions (10–13). These data are difficult to reconcile with the C_H gene deletion model because they imply that a V_H gene can be expressed with the C_δ gene without concomitant deletion of the C_μ gene. A number of mechanisms have been suggested to account for these observations, including the "copy-insertion" mechanism (14) in which a copy of the V_H gene is joined to the C_δ gene while the original remains joined to the C_μ gene, and differential RNA processing of a single transcript containing V_H, C_μ, and C_δ genes (15).

This paper reports experiments carried out to test these hypotheses. The results support RNA splicing as the mechanism by which the C_μ and C_δ genes are expressed simultaneously in B cells and also suggest that V_H-to-C_δ DNA switch recombination may occur in plasma cells that express only IgD.

MATERIALS AND METHODS

Rat IgD Myeloma. The rat myeloma IR731 (16), a plasmacytoma, was passaged subcutaneously in Lou/M/Wsl N rats (NIH). Total cell poly(A)$^+$RNA was prepared from IR731 by a method similar to that of Chirgwin *et al.* (17), followed by oligo(dT)-cellulose chromatography. This RNA was hydrolyzed with base to an estimated average size of ≈500 nucleotides and labeled with ^{32}P by using [γ-^{32}P]ATP and polynucleotide kinase for use as a probe. Cytoplasmic poly(A)$^+$RNA for RNA blots was prepared from IR731 as described (18).

BALB/c $\mu^+\delta^+$ Lymphomas. These lymphomas, the generous gifts of R. Asofsky and K. Jin Kim, were passaged subcutaneously in BALB/c mice (19). The presence of cell-surface IgM and IgD molecules was verified by immunofluorescence using anti-IgM (Cappel) and monoclonal anti-IgD (Becton Dickinson) reagents. GCL-2.1 cells were from W. Raschke. DNA from these and other tissues was prepared by the method of Blin and Stafford (20).

Germ-Line C_δ Clones. The germ-line clone ChSpμ7 has been described (4, 21). The clone ChSp37 was isolated from the same library of mouse sperm DNA as ChSpμ7. The positions of C_δ sequences in these clones were determined by hybridization of [5′-^{32}P]poly(A)$^+$RNA from IR731 to blots of restriction digests of these clones. Subclones of the hybridizing re-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviations: V_H, heavy chain variable; C_H, heavy chain constant; kb, kilobase(s).

gions were generated by ligation of restriction fragments into the corresponding site or sites of pBR322 and were used as probes in Southern (22) and RNA blot experiments. The cDNA clone p104Eμ12 (21) was used as a probe for C_μ sequences. The J_H probe, containing the J_H gene cluster and 3′-flanking sequence, was prepared by M. Steinmetz.

BALB/c $\mu^+\delta^+$ Normal B-Cell DNA Library. IgM-positive cells were isolated from BALB/c spleens by using a fluorescence-activated cell sorter. Ninety-three to 97% of the purified cell population stained positively for surface IgM, and 99% was positive for surface IgD. A library of 12×10^6 recombinant phage was constructed by ligation of *Eco*RI partial digests of $\mu^+\delta^+$ spleen cell DNA to phage vector Charon 4A (23), followed by *in vitro* packaging (24).

All manipulations of microorganisms containing recombinant DNA were carried out under P2/EK2 or P2/EK1 conditions prior to January 1980, after which P1/EK2 and P1/EK1 conditions were used.

DNA sequence analysis was as described (25).

RESULTS AND DISCUSSION

Characterization of Genomic Clones. [5′-^{32}P]Poly(A)$^+$RNA prepared from IR731 was used to screen a number of genomic clones known to contain mouse immunoglobulin C_H region genes or their flanking sequences. Two BALB/c sperm DNA clones, ChSpμ7 and ChSp37, hybridized to this probe. These clones were subjected to restriction map analysis (Fig. 1). The [5′-^{32}P]poly(A)$^+$RNA probe hybridized to three discrete regions of the cloned DNA.

Identification of the C_δ Gene. Restriction fragments hybridizing to IR731 poly(A)$^+$RNA were subcloned into the plasmid vector pBR322 (Fig. 1). A DNA sequence determined near the 3′ end of the pδ24 clone (Fig. 2, part B) appeared to encode the first third of an immunoglobulin domain, with a cysteine residue and several conserved amino acids in the appropriate positions. This amino acid sequence is translated from the only open reading frame and is associated with a possible downstream splicing site. This sequence displays striking homology to the protein sequence of the $C_\delta 3$ domain of human IgD (26): 18 of 33 residues, including a stretch of 10 surrounding the

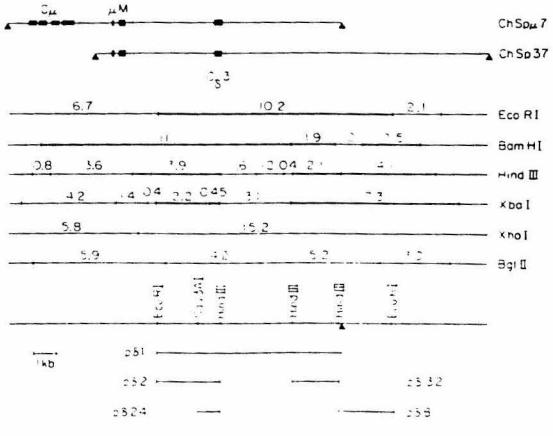


FIG. 1. Restriction map of ChSpμ7 and ChSp37. Restriction fragments of a particular digest that hybridize to [5′-^{32}P]poly(A)$^+$RNA from IR731 are indicated by heavy lines. *Bgl* II and *Xho* I digests were not tested for hybridization to IR731 RNA. Plasmid subclones containing hybridizing regions are also indicated. μM, IgM membrane exons; ▲, synthetic *Eco*RI linker site.

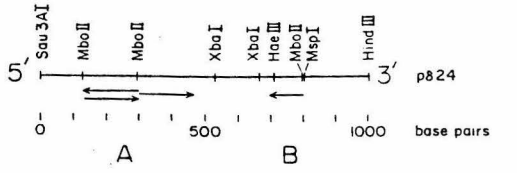


FIG. 2. Partial restriction map of pδ24 and strategy for determination of nucleotide sequence.

cysteine residue are identical, and an additional 10 are either conservative substitutions or can be attributed to a single-base difference in the genetic code. Thus, this sequence codes for the $C_\delta 3$ domain of mouse IgD.

Additional sequence determinations in the 5′ region of pδ24 (Fig. 2, part A) revealed no additional immunoglobulin domain coding sequences, and the unsequenced portion (coordinates 470–700) is too small to encode a complete immunoglobulin domain. Additional domain and hinge regions presumably lie in pδ2 because no hybridization of IR731 RNA was observed 5′ to the 10.2-kb *Eco*RI fragment (Fig. 1). We conclude that in the germ-line the C_δ gene is separated from the C_μ M exons by >2 kb.

***C_δ* Gene Encodes Two mRNAs with Alternative 3′ Ends.** The germ-line DNA subcloned in pδ2 contains the $C_\delta 3$ domain of the δ gene and, as argued above, probably additional C_δ domains as well. Labeled total cellular poly(A)$^+$RNA from IR731 also hybridized with sequences 3–7 kb to the $C_\delta 3$ encoding sequence (Fig. 1). These regions were subcloned as pδ32 and pδ8. Both pδ32 and pδ8 contain only single-copy sequences and do not cross hybridize with each other or with pδ2. That this DNA also contains δ gene sequences was demonstrated by RNA blots with IR731 mRNA and nick-translated pδ32 and pδ8 as probes. Hybridization of C_δ probes (pδ2) to IR731 poly(A)$^+$mRNA revealed a major δ mRNA species of 1.6 kb and a minor species of 2.7 kb (Fig. 3). In addition, the 1.6-kb species hybridized with pδ32, and the 2.7-kb mRNA hybridized with pδ8. These results suggest that pδ32 and pδ8 contain separate, noncontiguous gene segments which represent alternative 3′ ends for the 1.6- and 2.7-kb mRNAs. At present we do not know whether each gene sequence is composed of one or more exons. Nonetheless, these results indicate that C_δ gene sequences that are contiguous in δ mRNA occupy about 8 kb of germ-line DNA.

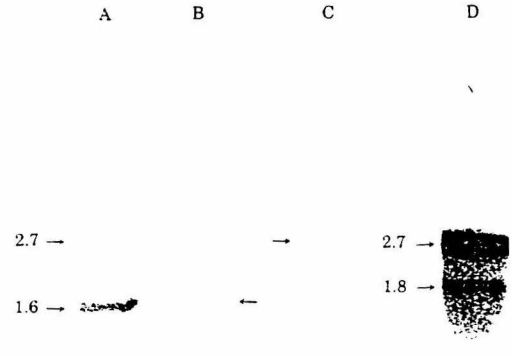


FIG. 3. RNA blots of cytoplasmic mRNA from IR731 rat myeloma (lanes A–C) and GCL-2.1 (lane D), a $\mu^-\delta^-$ mouse cell line, with nick-translated C_δ probes. Lanes: A, pδ2; B, pδ32; C, pδ8; Dpδ2. Fragment sizes are shown in kb.

We have also observed two corresponding species in a mouse $\mu^+\delta^+$ cell line (GCL-2.1) (Fig. 3).

Because the 1.6-kb mRNA is the major species in IR731 myeloma cells which secrete IgD, we propose that pδ32 encodes a δ_s terminus for δ chain secretion. The pδ8 sequence in the 2.7-kb mRNA may encode a δ_m terminus for membrane-bound δ chains. We propose that these alternative 3′ sequences are spliced to the $C_\delta 3$ domain to generate either δ_s or δ_m mRNA. This arrangement is different from that of μ_s and μ_m mRNA, in which the μ_s terminus encoding sequence is contiguous with the $C_\mu 4$ domain (15, 18).

$\mu^+\delta^+$ B-Cell DNA Contains Rearranged C_μ and Germ-Line C_δ Genes. The lymphoma lines L10A and K46 are B-cell tumors and express both IgM and IgD molecules on the cell surface (19). We carried out Southern blot experiments with DNA prepared from these tumor cells, using BALB/c embryo DNA as a control. Fig. 4A shows a Southern blot of a *Kpn* I digest of mouse embryo and K46 DNA with the C_μ cDNA clone as probe. In embryo DNA, this C_μ probe hybridizes to a fragment ≈13 kb long. This fragment contains both the C_μ gene and the J_H gene cluster (27). In K46 DNA, the C_μ gene appears on an ≈20-kb restriction fragment, which indicates that a DNA rearrangement has occurred, presumably the V-J joining event (27) (similar results were obtained with L10A DNA). Note that a faint band is present in K46 DNA which corresponds to the germ-line C_μ gene. Both K46 and L10A are polyploid (28), and this band could represent an unrearranged chromosome. Alternatively, it could arise from DNA derived from contaminating host tissue.

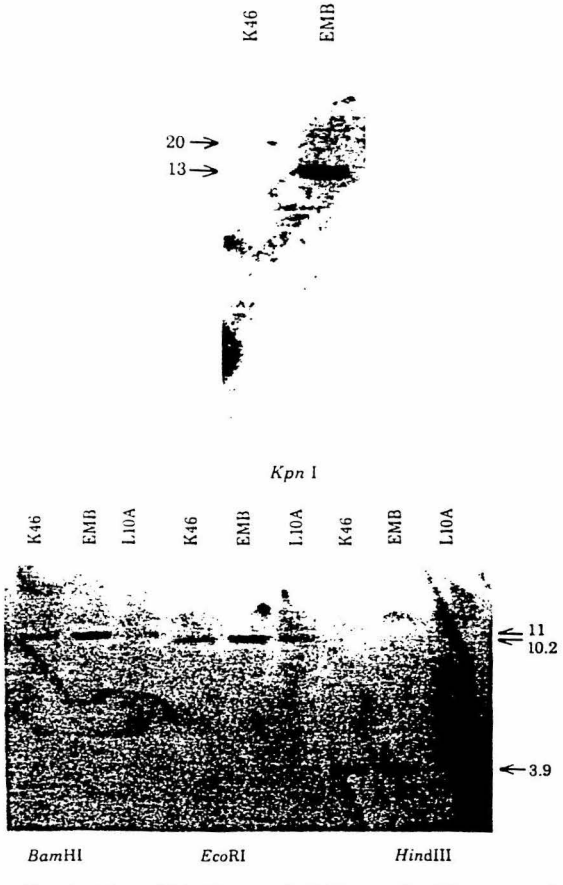


FIG. 4. (*Upper*) Hybridization of p104Eμ12 to *Kpn* I digests of K46 and mouse embryo DNA. (*Lower*) Hybridization of the $C_\delta 3$ probe to *Bam*HI, *Eco*RI, and *Hind*III digests of K46, mouse embryo, and L10A DNA. Hybridization conditions: 5X SET, 2X Denhardt's, 0.5% $NaDodSO_4$, 68°C.

The corresponding data for the C_δ gene are shown in Fig. 4*B* for K46, embryo, and L10A DNAs. The probe used in this experiment was a 0.68-kb *Bam*HI/*Xba* I restriction fragment from pδ2 which contains the $C_\delta 3$ domain and 340 nucleotides of pBR322. This C_δ probe hybridizes to identical bands in both embryo and lymphoma DNA restriction digests. No other hybridization could be detected. Furthermore, the bands observed in Fig. 4*B* are identical in size to those observed in ChSpμ7 and ChSp37, which demonstrates that these latter clones contain no detectable cloning artifacts.

Analogous experiments were carried out with pδ8 and pδ32 as probes. In each case, hybridization to germ-line and lymphoma DNA restriction digests gave similar results (data not shown), indicating that no rearrangement of the C_δ gene or its flanking sequences occurs in $\mu^+\delta^+$ lymphoma DNA.

We also have obtained similar results by using the $\mu^+\delta^+$ normal B-cell library. When 1.75×10^6 and 3.2×10^6 recombinant phage were screened with the C_δ and C_μ probes, respectively, a total of 4 C_δ clones and 11 C_μ clones were isolated. Six of the 11 C_μ clones also contained the J_H locus, and 5 of these showed rearrangements consistent with V_H-D-J_H joining. Restriction map analysis of the C_δ and C_μ clones revealed that they all contain only germ-line DNA within the C_μ gene, between C_μ and the downstream *Eco*RI site (Fig. 1), and in the C_δ gene and 3′-flanking sequence. The possibility that we simply failed to clone or detect a rearranged C_δ gene cannot be ruled out; however, these results are consistent with those described above and suggest that, insofar as can be detected by gel electrophoresis, the C_δ gene in $\mu^+\delta^+$ B cell and lymphoma DNA is unrearranged.

Thus, we conclude that the simultaneous expression of IgM and IgD molecules on the B lymphocyte surface does not involve rearrangement of the C_δ gene or, by implication, a V_H-C_δ DNA switch recombination.

IR731 Myeloma DNA Contains a Rearranged C_δ Gene and Has Deleted the C_μ Gene. Using the μ cDNA clone and the restriction fragment encoding the $C_\delta 3$ domain as probes, we also examined the disposition of the C_μ and C_δ genes in IR731 myeloma DNA. DNA from Lou/M/Wsl N rat liver served as a control (Fig. 5). Both probes cross hybridized with the rat genes on a Southern blot of Lou/M/Wsl N liver DNA. Results obtained with the C_μ probe indicate that the C_μ gene has been largely deleted from the myeloma DNA (Fig. 5A). The corresponding experiment with the $C_\delta 3$ probe and three restriction endonucleases revealed that the C_δ gene had undergone DNA rearrangement.

These data indicate that the C_δ gene in this IgD-producing myeloma has been rearranged, unlike that in mouse $\mu^+\delta^+$ B cells. Several explanations for these contrasting results should be considered. First, IgD expression in the rat may be fundamentally different from that in the mouse, although we believe that there exists no other evidence to indicate that this might be the case. Second, these observations could be an artifact of the myeloma condition. For example, the significance of "abortive" rearrangements observed in mouse plasmacytomas (29) is not understood. Finally, these results could conceivably be attributed to a minor substrain difference between Lou/M/Wsl N, the tumor host, and Lou/C/Wsl (16), in which IR731 appeared.

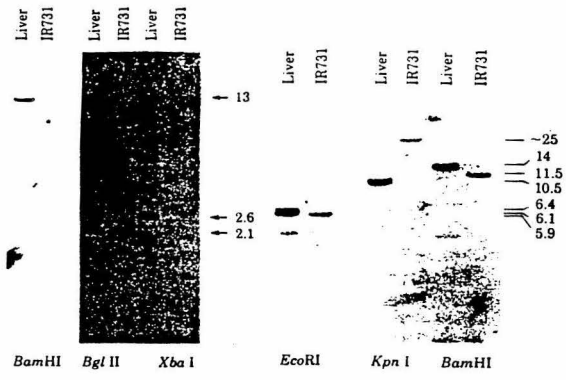


FIG. 5. (*Left*) Hybridization of p104Eμ12 to Lou/M/Wsl N liver and IR731 DNA digested with *Bam*HI, *Bgl* II, and *Xba* I. (*Right*) Southern blot of Lou/M/Wsl N liver and IR731 DNA restriction enzyme digests with the C_δ3 probe. (*C*) Hybridization of the J_H probe to *Hin*dIII and *Bam*HI digests of Lou/MN liver and IR731 DNA. Hybridization conditions were: 5X SET, 2X Denhardt's, 0.5% NaDodSO$_4$, 62°C. Fragment sizes are in kb.

To address the latter two questions, we carried out Southern blot experiments using pδ8, pδ32, a γ3 cDNA clone from myeloma J606, and a plasmid subclone containing the mouse J_H gene cluster and its 3′ flanking sequence. The pδ8 and pδ32 clones gave similar results with Lou/M/Wsl N liver and Lou/C/Wsl myeloma (IR731) DNA: in each case, both hybridized to a single 5.7-kb *Eco*RI fragment. The γ3 cDNA probe also hybridized to 9.2-, 14-, and 16-kb *Eco*RI fragments in both DNA samples. Thus, the C_γ genes and the counterparts of pδ32 and pδ8 in the rat genome display no polymorphism in the two substrains. These results suggest that DNA rearrangement rather than restriction enzyme site polymorphism is responsible for the observed difference in hybridization of the C_δ3 probe to Lou/M/Wsl N liver and IR731 DNA and that this rearrangement has occurred on the 5′ side of pδ32 and pδ8.

Although the C_μ gene had been deleted from IR731 DNA, results obtained with the J_H probe (Fig. 5*C*) indicate that the J_H genes and 3′-flanking sequence are still present and have been rearranged, as would be expected for a *V-D-J*$_H$ joining event. This observation suggests that a V_H–C_δ switch recombination has indeed occurred in IR731 DNA.

Two Alternative Mechanisms for the Expression of IgD. Based on the results described above, we propose the existence of two different molecular mechanisms for the expression of the C_δ gene. First, in $\mu^+\delta^+$ B cells, these results suggest regulation at the level of RNA processing, probably involving multiple sites for splicing and poly(A) addition. This type of control has been observed in late adenovirus mRNA processing (30–32) and has been implicated in the synthesis of membrane-bound and secreted IgM (μ_m and μ_s) mRNA from a single transcript (15). This mechanism presupposes the existence of a poly(A) addition site 3′ to the C_δ gene. Extension of a transcript beyond the poly(A) addition sites of μ_s and μ_m mRNA would then generate the precursor of δ mRNA. Wall *et al.* (33) pointed out that, according to this model, the preferential utilization of the μ_s poly(A) site which would accompany the initiation of μ_s synthesis would halt production of membrane-bound IgM and IgD molecules. Results have been obtained which agree with this prediction (34–36). The C_δ gene thus could be another example in eukaryotes in which developmentally regulated RNA splicing generates alternative protein forms. We believe that this control mechanism will be a general one in eukaryotic gene expression.

The second mechanism, apparently used by the rat IgD myeloma IR731, involves deletion of C_μ and rearrangement of C_δ, presumably via a V_H–C_δ switch recombination. We infer the presence of one or more switch sites (37), probably in the region between C_δ and the IgM M exons. At present, the mechanism of this particular switching event remains unknown. We predict that, generally, IgD-secreting cells also will exhibit rearrangement of the C_δ gene. A schematic representation of these mechanisms is given in Fig. 6.

C_H Gene Linkage Family and C_H Gene Expression. The data in this paper and those from other laboratories (38–40), allow construction of a linkage map of the immunoglobulin heavy chain gene family (Fig. 6, top line). Because the separation between the C_μ and C_δ genes is substantially smaller than that for other C_H genes, we believe that the C_μ–C_δ system is likely to be the only pair of C_H genes that uses an RNA processing mechanism as proposed above for their expression. Expression of other heavy chain genes (C_γ, C_α, C_ε) most likely will occur only by the C_H switching mechanism involving DNA rearrangement.

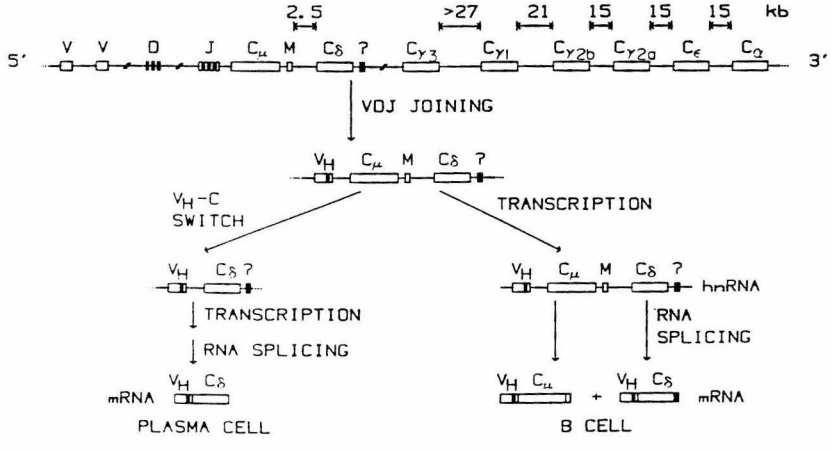


FIG. 6. Proposed mechanisms for expression of IgD in B cells and plasma cells. The top line represents the current germ-line linkage map (38–40) of the immunoglobulin C_H locus. Known distances are in kb. The putative δ_m exon is indicated by ?.

Note. After this manuscript was submitted for review, two articles (41, 42) describing detailed structural studies of the mouse C_δ gene appeared. Our data on the structure of the C_δ gene largely agree with the data presented in them.

The authors thank Dr. B. Clevinger and Dr. J. M. Davie for their invaluable help in obtaining and propagating IR731, Dr. K. Jin Kim and Dr. R. Asofsky for their gifts of K46 and L10A, and Dr. W. Raschke for GCL-2.1. We are grateful to Dr. F. Putnam for making his amino acid sequence of the human δ chain available prior to publication. We also thank Mr. B. Granger and Dr. E. Lazarides for use of the fluorescence microscope. This research was supported by Grant AI 16913 from the National Institutes of Health to L.E.H. and Grant PCM 7924876 from the National Science Foundation to R.W. K.W.M. is a Fellow in Cancer Research supported by Grant DRG-313-FT of the Damon Runyon–Walter Winchell Cancer Fund.

1. Godin, J. W. (1978) *Cont. Topics Immunobiol.* 8, 203–243.
2. Early, P., Huang, H., Davis, M., Calame, K. & Hood, L. (1980) *Cell* 19, 981–992.
3. Sakano, H., Maki, R., Kurosawa, Y., Roeder, W. & Tonegawa, S. (1980) *Nature (London)* 286, 676–683.
4. Davis, M. M., Calame, K., Early, P. W., Livant, D. L., Joho, R., Weissman, I. & Hood, L. (1980) *Nature (London)* 283, 733–739.
5. Maki, R., Traunecker, A., Sakano, H., Roeder, W. & Tonegawa, S. (1980) *Proc. Natl. Acad. Sci. USA* 77, 2138–2142.
6. Kataoka, T., Kawakami, T., Takahashi, N. & Honjo, T. (1980) *Proc. Natl. Acad. Sci. USA* 77, 919–923.
7. Honjo, T. & Kataoka, T. (1978) *Proc. Natl. Acad. Sci. USA* 75, 2140–2144.
8. Yaoita, Y. & Honjo, T. (1980) *Nature (London)* 286, 850–853.
9. Herzenberg, L. A., Herzenberg, L. A., Black, S. J., Loken, M. R., Okumura, K., van der Loo, W., Osborne, B. A., Hewgill, D., Goding, J. W., Gutman, G. & Warner, N. L. (1976) *Cold Spring Harbor Symp. Quant. Biol.* 41, 33–45.
10. Salsano, F., Froland, S., Natvig, J. B. & Michaelsen, T. E. (1974) *Scand. J. Immunol.* 3, 841–846.
11. Fu, S. M., Winchester, R. J. & Kunkel, H. G. (1975) *J. Immunol.* 114, 250–252.
12. Goding, J. W. & Layton, J. E. (1976) *J. Exp. Med.* 144, 852–857.
13. Stern, C. & McConnell, I. (1976) *Eur. J. Immunol.* 6, 225–227.
14. Williamson, A. R. (1976) *Annu. Rev. Biochem.* 45, 467–500.
15. Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R. & Hood, L. (1980) *Cell* 20, 313–319.
16. Bazin, H., Beckers, A., Urbain-Vansanten, G., Pauwels, R., Bruyns, C., Tilkin, A. F., Platteau, B. & Urbain, J. (1978) *J. Immunol.* 121, 2077–2082.
17. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. & Rutter, W. J. (1979) *Biochemistry* 18, 5294–5299.
18. Rogers, J., Early, P., Carter, C., Calame, K., Bond, M., Hood, L. & Wall, R. (1980) *Cell* 20, 303–312.
19. McKeever, P. E., Kim, K. J., Nero, G. B., Laskov, R., Merwin, R. M., Logan, W. J. & Asofsky, R. (1979) *J. Immunol.* 122, 1261–1265.
20. Blin, N. & Stafford, D. W. (1976) *Nucleic Acids Res.* 3, 2303–2308.
21. Calame, K., Rogers, J., Early, P., Davis, M., Livant, D., Wall, R. & Hood, L. (1980) *Nature (London)* 284, 452–455.
22. Southern, E. M. (1975) *J. Mol. Biol.* 98, 503–517.
23. Blattner, F. R., Williams, B. G., Blechl, A. E., Denniston-Thompson, K., Faber, M. E., Furlong, L.-A., Grunwald, D. J., Kiefer, D. O., Moore, D. D., Schumm, J. W., Sheldon, E. L. & Smithies, O. (1977) *Science* 196, 161–169.
24. Hohn, B. & Murray, K. (1977) *Proc. Natl. Acad. Sci. USA* 74, 3259–3263.
25. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* 74, 560–564.
26. Lin, L.-C. & Putnam, F. (1981) *Proc. Natl. Acad. Sci. USA* 78, 504–508.
27. Forster, A., Hobart, M., Hengartner, H. & Rabbits, T. H. (1980) *Nature (London)* 286, 897–899.
28. Kim, K. J., Kanellopoulos-Langevin, C., Merwin, R. M., Sachs, D. H. & Asofsky, R. (1979) *J. Immunol.* 122, 549–554.
29. Coleclough, C., Cooper, D. & Perry, R. P. (1980) *Proc. Natl. Acad. Sci. USA* 77, 1422–1426.
30. Nevins, J. R. & Darnell, J. E. (1978) *J. Virol.* 25, 811–823.
31. Ziff, E. & Fraser, N. (1978) *J. Virol.* 25, 897–906.
32. Darnell, J. E. (1979) *Prog. Nucl. Acid Res. Mol. Biol.* 22, 327–353.
33. Wall, R., Choi, E., Carter, C., Kuehl, M. & Rogers, J. (1980) *Cold Spring Harbor Symp. Quant. Biol.* 45, in press.
34. Bourgois, A., Kitajima, K., Hunter, I. R. & Askones, B. A. (1977) *Eur. J. Immunol.* 7, 151–153.
35. Kearney, J. F. & Abney, E. R. (1978) *Cont. Topics Immunobiol.* 8, 245–265.
36. Preudhomme, J. L. (1977) *Eur. J. Immunol.* 7, 191–193.
37. Davis, M., Kim, S. & Hood, L. (1980) *Science* 209, 1360–1365.
38. Nishida, Y., Kataoka, T., Ishida, N., Nakai, S., Kishimoto, T., Bottcher, I. & Honjo, T. (1981) *Proc. Natl. Acad. Sci. USA* 78, 1581–1585.
39. Shimizu, A., Takahashi, N., Yamawaki-Kataoka, Y., Nishida, Y., Kataoka, T. & Honjo, T. (1981) *Nature (London)* 289, 149–153.
40. Honjo, T., Nishida, Y., Shimizu, A., Takashi, N., Kataoka, T., Obata, M., Yamawaki-Kataoka, T., Nikaido, T., Nakai, S., Yaoita, Y. & Ishida, N. (1981) *Proceedings of Symposium on Mechanisms in Mucosal Immunity*, eds. Hanson, L. A., Sell, K. W. & Strober, W. (Raven, New York), in press.
41. Liu, C. P., Tucker, P. W., Mushinski, J. F. & Blattner, F. (1980) *Science* 209, 1348–1353.
42. Tucker, P. W., Liu, C. P., Mushinski, J. F. & Blattner, F. (1980) *Science* 209, 1353–1360.

VII. SUMMARY

The work described in this thesis is meant to reflect not only the diversity of the efforts I have been involved with during my tenure at Caltech, but to indicate also the unifying themes of these efforts. The intent of this summary is to provide some direction from which to evaluate the impact this work has had on the fields represented here. Most of my biological contributions have involved the direct investigation of the genes and proteins of the immunoglobulin gene superfamily. Many of these papers provided the first analytical description and details of the genes encoding various members of the superfamily as well as their initial identification (e.g., Myelin Associated Glycoprotein - MAG). Each paper provided a synthesis of the experimental detail into a functional and evolutionary context. Of particular importance have been the novel descriptions provided concerning the structural motifs of the superfamily and the evolutionary implications of the diversifying potential of these motifs. Coincident with one other author, this work was the first to describe an entire new class of immunoglobulin-related motifs. This sequence represents a unique, probably more nearly primordial view of the sequence from which the rest of the superfamily arose. It is clear from this view that immune recognition was not the driving force in the development of the immunoglobulin homology unit. Rather, it was the inherent diversifying potential of the homology that drove the development of several complex receptor/ligand systems including the immune system.

The most unifying theme of these efforts has concerned the diversity potential of the immunoglobulin motif and its functional and evolutionary significance. The work represented here contributed directly to the discovery of several of the somatic diversifying mechanisms of the rearranging immune receptor genes, including junctional joining diversity and the first discovery of a eukaryote coding element that is routinely translated in all three frames. It is clear from these analyses that the number of potential specificities arising as a result of these and other mechanisms far exceeds the number of immune cells ever generated in any one animal. These discoveries have helped shaped our current view of how the limited number of genetic elements of the immune receptor gene families can accommodate and recognize the universe of potential antigens.

The analyses of the diversity of the T-cell receptor proteins were the most comprehensive for their time and have led directly to models of immune function by myself and others. Our analyses of the structural potential of T-cell receptors lead to one of the first thorough models describing the likely similarity of the T-cell receptor and immunoglobulin structures, a view widely held today but not so at the time. Again, these models have directly impacted our view of how these molecules function and have lead to much further experimentation in this and other laboratories.

The included structural and functional investigations have generated specific evolutionary models involving the role of immunoglobulin motif. This motif may be unique in its ability to accommodate sequence diversity while maintaining its complex structural definition. This is no doubt responsible for its use as the main structural component of an incredibly diverse array of functionally unrelated molecules, involved in everything from cartilage formation to fine-specificity antigen recognition. This collection represents the most comprehensive discussions on the implications this potential and other properties of the motif have with regards to the evolution of complex eukaryotic traits. Particularly, these discussions focus on the role of relatively saltational changes and canalization in evolution.

Although not explicitly included in this collection, many of the contributions of the included work resulted from unique technologies and tools developed to support the specific biological investigations. The software tools created to carry out much of the analyses in this thesis have also been used and cited by dozens of other biologists. Algorithms developed have found their way into commercial software packages available to the entire community.

In summary, most of my work at Caltech has been collaborative and generally of a theoretical or analytical nature. Thus, I have included many papers for which I am not the principal author. However, in each of these papers my contributions have been explicit and significant to understanding the data generated and, in many, to the experimental direction itself. As diverse as these efforts may seem, they have all contributed to my goal of investigating and understanding the underlying principals of evolution as represented by the development of complex systems. My future efforts in molecular biology will continue to follow this model as it provides me the means by which to effectively work on the many diverse systems required to maintain the perspective necessary for further insight into the co-evolutionary issues concerning the complex interactions of the immunoglobulin gene superfamily. I expect an explosion of information to result from the further refinement and use of the instrumentation and computer technology for which I have been involved in the development. I hope to be able to further contribute to the thorough understanding of this new knowledge.

The work included here has contributed to not only a large amount of descriptive detail concerning the immunoglobulin gene superfamily, but also to many testable models concerning both specific and general aspects of the function and evolution of molecules that contain the

immunoglobulin motif. At the same time, I have also been involved in the development of some of the most exciting technologies of modern molecular biotechnology. My future goals do not differ significantly from these efforts. It is clear from the papers included in this collection, that an abiding interest in the biology of a fascinating gene family has been and will continue to be the driving force of my efforts.

VIII. APPENDIX

CONTRIBUTIONS TO EACH PAPER

Diversity of the Immunoglobulin Gene Superfamily

Tim Hunkapiller and Leroy Hood

This paper was the introductory chapter for Advances in Immunology for 1989. It is a comprehensive review of the immunoglobulin gene superfamily as well as a significant synthesis of the knowledge regarding this family. It was written by me with editorial help from Lee Hood. There are many observations and models included in the paper that are novel to the literature, particularly as regards diversification potential of the rearranging immune receptor genes and structural distinctions between superfamily classes. The paper was specifically written to act as an introduction and summation of much of my own research represented more specifically in the preceeding papers. I have researched and published on members of essentially every class of immune-related receptors of the superfamily in an effort to better understand both their evolutionary and functional relationships.

A Speculative View of the Multicomponent Nature of T Cell Antigen Recognition

Joan Goverman, **Tim Hunkapiller** and Leroy Hood

This paper was a completely collaborative effort between myself and Joan Goverman. It presents a synthesis of the emerging molecular and functional data concerning antigen recognition and cellular activation of the regulator cells of the immune response, T cells. A detailed model is presented that substantially diverged from much of the contemporary dogma. It is presented as part of the introduction to the thesis to present the functional implications of many of the observations I have been associated with, particularly as regards the likely structure and diversification of the immune receptors. It is our contention that the general tenets of the model are still valid given our current state of knowledge of the processes discussed.

The Impact of Modern Genetics on Evolutionary Theory

Tim Hunkapiller, Henry Huang, Leroy Hood and John Campbell

This is the first paper for which I was the primary author. It is a chapter in the book Perspectives on Evolution. It is a discussion on the emerging view of molecular phenomena in the context of evolutionary thought. The major theme revolved around models of rapid, punctuated phenotypic change facilitated by only limited genetic change. We use the multigene families of the immune receptors and models of their evolution as particularly important examples in the discussion.

T Cell Antigen Receptors and the Immunoglobulin Supergene Family

Leroy Hood, Mitchell Kronenberg and **Tim Hunkapiller**

This was the first major paper we published specifically on the immunoglobulin gene superfamily as the complexity of the relationships of its member genes was beginning to become particularly apparent. It was an invited review that resulted specifically from work we had done on characterizing the T-cell immune receptor and genes of the major histocompatibility complex. The writing was a shared process. It contains the first proposed genealogy for the superfamily members known at that time.

The Growing Immunoglobulin Gene Superfamily

Tim Hunkapiller and Leroy Hood

This was an invited News and Views for *Nature* written principally by me. It was prompted by observations that superfamily members were likely to be prominent in non-immune systems such as the nervous system. The suggestion was first made here that the immunoglobulin-like domains of neural cell adhesion molecule (N-CAM) were responsible for the homotypic interactions of these molecules. This is generally accepted now and has had implications for other, subsequently described members of the superfamily. Also, this is the first report that another neural protein (now known to be myelin associated glycoprotein [MAG]) was also a member of the superfamily and had an analogous structure to N-CAM.

L3T4 and the Immunoglobulin Gene Superfamily: New Relationships Between the Immune System and the Nervous System

Jane R. Parnes and **Tim Hunkapiller**

This is a review and synthesis of the molecules of the immunoglobulin gene superfamily that are expressed in the nervous system. L3T4 (on which we had published earlier) was the focus of the article due to its expression in both the nervous and immune systems. The writing was shared. I particularly was responsible for the analysis and discussion of the sequence data and the evolutionary discussion.

Implications of the Diversity of the Immunoglobulin Gene Superfamily

Tim Hunkapiller, Joan Goverman, Ben Koop and Leroy Hood

This is an extension of the *Advances in Immunology* paper that introduces the thesis. There is significantly more discussion regarding the structural paradigms represented in the superfamily as well as a more detailed view of its evolutionary history. The evolutionary implications of the diversity potential of the immunoglobulin gene superfamily motif is the primary focus of the paper. I was the primary author of the text and models presented. Particularly, a novel model for the evolution and original function of immune receptor rearrangement and allelic exclusion is presented.

Mouse T Cell Antigen Receptor: Structure and Organization of Constant and Joining Gene Segments Encoding the beta Polypeptide

Marie Malissen, Karyl Minard, Shelly Mjolsness, Mitchell Kronenberg, Joan Goverman, **Tim Hunkapiller**, Michael Prystowsky, Yasunobu Yoshikai, Frank Fitch, Tak Mak and Leroy Hood

This was the first paper to describe the genomic organization of one of the two chains of the T-cell receptor. I did not participate in generation of the data presented. However, I participated in the evaluation, interpretation and discussion of the non-sequence data and was essentially solely responsible for the analysis and discussion of the sequence information. Only the first author contributed more to the text of the paper than myself.

Rearranged beta T Cell Receptor Genes in a Helper T Cell Clone Specific for Lysozyme: No Correlation Between V beta and MHC Restriction

Joan Goverman, Karyl Minard, Nilabh Shastri, **Tim Hunkapiller**, Dan Hansburg, Eli Sercarz and Leroy Hood

This paper was the first major discussion concerning the diversity of the T-cell receptor. Its most striking observation was that a single beta V gene can be used for receptors for different MHC specificities. This had significant implications subsequently for the models of MHC restriction and T-cell receptor diversity. I participated fully in the evaluation of all forms of data and much of the experimental design. I was responsible for all sequence analyses. This led to the discovery of a novel diversifying mechanism for immune receptors. Only the first author contributed more to the text than myself. I contributed fully to the discussions of the biological significance of the data.

The Murine T-Cell Receptor Uses a Limited Repertoire of Expressed V beta Gene Segments

Richard Barth, Byung Kim, Nancy Lan, **Tim Hunkapiller**, Nancy Sobieck, Astar Winoto, Howard Gershenfeld, Craig Okada, Dan Hansburg, Irving Weissman and Leroy Hood

At the time, this was the most thorough analysis of the diversity expressed in the beta chain of the T-cell receptor. The paper was at odds with several contemporary models and analyses, but has proven correct (or at least not incorrect) in almost all particulars. Lee Hood and myself were primarily responsible for the final form of the text. I was solely responsible for the sequence analyses and, with Leroy Hood, for most of the biological discussion. I developed several software tools to facilitate the sequence analyses and employed a novel modification to secondary structure prediction tools to explore the similarity of T-cell and immunoglobulin structures.

Chimeric Immunoglobulin-T Cell Receptor Proteins Form Functional Receptors: Implications for T Cell Receptor Complex Formation

Joan Goverman, Stephen Gomez, Kathleen Segesman, **Tim Hunkapiller**, Walter Laug and Leroy Hood

This paper describes the results of activation experiments using molecules encoded by chimeric genes constructed with an immunoglobulin variable region gene and one of two T-cell constant region genes. We were able to indicate a fundamental similarity between the functional structures of immunoglobulins and T-cell receptors as well as provide insight into possible mechanisms of T-cell receptor function. I designed the chimeric molecules for which the chimeric genes were constructed from analyses of the structural and functional properties of the two receptor molecules. The biological discussion of the paper was a joint effort of Joan Goverman and myself. I contributed most of the structural discussion. I presented a novel model of the T-cell receptor as possibly a stretch-activated receptor analogous to stretch-activated ion channels. We also questioned certain dogma concerning the structure of the T-cell alpha chain. This paper is one effort to address issues discussed in the functional model paper that introduces this section.

The T Cell Differentiation Antigen Leu-2/T8 Is Homologous to Immunoglobulin and T Cell Receptor Variable Regions

Vikas P. Sukhatme, Kurt Sizer, Amy Vollmer, **Tim Hunkapiller** and Jane Parnes

This paper is the first molecular description of one of the so-called T-cell accessory molecules and its characterization as a member of the immunoglobulin gene superfamily. These cell-surface antigens facilitate T-cell antigen binding and activation and play a significant role in MHC restriction phenomena. I was responsible for all sequence analyses and software. The senior author and myself were responsible for the structural and evolutionary discussion of the paper. This is the first of a series of collaborations between myself and the senior author.

Isolation and Sequence of the L3T4 Complementary DNA Clones: Expression in T Cells and Brain.

Beatrice Tourvielle, Scott Gorman, Elizabeth Field, **Tim Hunkapiller** and Jane Parnes

This is the first molecular description of the mouse homologue of the human CD4 T-cell accessory molecule, which, along with CD8, is primarily responsible for MHC class restriction. I did all sequence and structural analyses and contributed to the biological discussion. Direct and comparative analyses of the sequence led to several novel observations.

Three cDNA Clones Encoding Mouse Transplantation Antigens: Homology to Immunoglobulin Genes

Michael Steinmetz, John Frelinger, Douglass Fisher, **Tim Hunkapiller**, Dennis Pereira, Sherman Weissman, Hiroshi Uehara, Stanley Nathenson and Leroy Hood

This paper reports the first significant DNA sequence of transplantation antigen cDNAs. I developed a novel sequence comparison software tool and statistical method to test the hypothesis that transplantation antigen genes were evolutionarily homologous to immunoglobulin genes. I contributed the discussion on the sequence data and their evolutionary history.

DNA Sequence of the Gene Encoding the E alpha Ia Polypeptide of the BALB/c Mouse

Janet McNicholas, Michael Steinmetz, **Tim Hunkapiller**, Patricia Jones and Leroy Hood

This is the first report of the genomic sequence of the gene encoding the alpha chain of the mouse I-E immune response antigen (Ia). I contributed the discussion on the evolutionary possibilities and sequence comparisons and did all of the sequence analyses.

Nucleotide Sequence of a Light Chain Gene of the Mouse I-A Subregion: A beta^d

Marie Malissen, **Tim Hunkapiller** and Leroy Hood

This is the first description of the genomic sequence of the gene for the beta chain of the mouse I-A immune response antigen (Ia). Marie Malissen and I wrote the paper. I am responsible for all sequence analysis and structural and evolutionary discussion. I developed novel software that aided in the prediction of an exon not found in the cDNAs of known human homologues. I read and, with the first author, aided in confirming all base calls from the raw sequence data.

Comparison of Exon 5 Sequences from 35 Class I Genes of the BALB/c Mouse

Kurt Brorson, Stephen Hunt III, **Tim Hunkapiller**, Henry Sun, Hilde Cheroutre, Deborah Nikerson and Leroy Hood

This paper compares a particular exon from each of the known mouse class I MHC genes to better understand the evolutionary and functional aspects of this group of sequences. I supervised the first author in the analyses of the data and the writing of the paper as well as the completion of the data acquisition.

The Joining of V and J Gene Segments Creates Antibody Diversity

Martin Wiegert, Robert Perry, Dawn Kelly, **Tim Hunkapiller**, James Schilling and Leroy Hood

This paper is the first to show that junctional diversity of the kappa immunoglobulin light chain probably results from imprecise joining of the variable and joining gene segments during rearrangement. This is my first paper from this lab. I generated all of the new protein sequence data presented in the paper, starting from raw mouse ascites.

Expression of IgD May Use Both DNA Rearrangement and RNA Splicing Mechanisms

Kevin Moore, John Rogers, **Tim Hunkapiller**, Phil Early, Carol Nottenburg, Irving Weissman, H. Bazin, Randy Wall and Leroy Hood

This paper is the first to describe how the same variable region can be present in different antibodies on a cell surface simultaneously with two different constant regions, mu and delta. The work done at Caltech was essentially done by Kevin Moore and myself. I participated in all facets of the experimental work; including tumor amplification, immunofluorescent cellular assays, and RNA and DNA blot analyses.