LINKAGE DISEQUILIBRIUM BETWEEN DNA POLYMORPHISMS IN A NATURAL POPULATION OF *Drosophila melanogaster* Meigen

JAMES NEIL MACPHERSON

Ph.D
UNIVERSITY OF EDINBURGH
1989

I declare that the work presented here is original and my own.

James N. Macpherson
October 1989

ABSTRACT

Samples of 44 extracted X chromosomes and 72 extracted third-chromosomes from a North Carolina natural population of *Drosophila melanogaster* Meigen were analysed for restriction-map variation at two loci, the *achaete-scute* complex on the X chromosome and the *rosy-Ace* region on chromosome III. A set of enzymes recognizing four-base-pair sequences was used in conjunction with a series of cloned probes homologous to the two regions with the aim of detecting linkage disequilibrium between polymorphic sites separated by known genetic and molecular distances.

In the *ac-sc* complex, extensive pairwise and three-way disequilibrium was found between most sites including those as far apart as 86 kb, spanning a tract of DNA encompassing several genes. Little or no tendency for disequilibrium to decline with increasing separation of variants was apparent. As crossing-over is virtually absent in this region, any disequilibrium arising either by mutation and drift or by natural selection and genetic hitchhiking would not be rapidly disseminated. In contrast, a rapid decay of disequilibrium with distance was observed in the vicinity of the *rosy* gene, no appreciable disequilibrium being detected for pairs of variants more than 0.5 kb apart. This pattern has been observed for some other *Drosophila* genes, most notably the *white* locus.

Overall heterozygosity was estimated at θ=0.001 4 in the *ac-sc* complex and θ=0.007 5 in the *ry* region, suggesting a lower level of nucleotide variability at *ac-sc* compared to other *Drosophila* loci although statistically the reduction is not significant. Variation at the DNA level would appear to harbour important information concerning the evolutionary forces acting on particular loci.

TABLE OF CONTENTS

1 INTRODUCTION 1

1.1 Linkage disequilibrium 1
1.2 Molecular variation 7
1.3 Aim of the survey 9
1.4 Genetic natural history of the regions surveyed 10
1.5 Strategy for surveying the regions 14

2 METHODS 16

2.1 Bacterial stocks 16
2.2 Bacteriophage stocks 16
2.3 *In vitro* packaging of bacteriophage DNA 17
2.4 Large-scale preparation of bacteriophage DNA 17
2.5 Marked *Drosophila* stocks 18
2.6 Chromosome extraction procedure 19
2.7 Purification of loci 22
2.8 Cytotype testing 23
2.9 Preparation of genomic DNA 24
2.10 Digestion of DNA with restriction endonuclease 25
2.11 Agarose gel electrophoresis and transfer 25
2.12 *poly*Propenamide gel electrophoresis and transfer 26
2.13 Ligation and subcloning 27
2.14 Transformation 28
2.15 Preparation of plasmid DNA 29
2.16 Synthesis of radioactively-labelled probes 30
2.17 Hybridization of filters to radioactively-labelled probes 31
2.18 Removal of probe DNA from membranes 31
2.19 DNA used as probes 32
2.20 Analysis of genomic DNA 34

3 RESULTS 36

3.1 The *ac-sc* region 36
3.1.1 M55 polymorphism 37
3.1.2 C67 polymorphism 39
3.1.3 H28 polymorphism 39
3.1.4 Six-cutter polymorphisms and insertion IV 39
3.2 The *ry-Ace* region 46
3.2.1 H-167 polymorphism 46
3.2.2 A-167 polymorphisms 48
3.2.3 *Cfo*I polymorphisms: fine-scale variation 48
3.2.4 Six-cutter polymorphisms 52
3.2.5 Extension across the region 55

4 ANALYSIS OF RESULTS 58

4.1 Measures of linkage disequilibrium 58
4.1.1 Two loci 58
4.1.2 Three loci 59
4.2 Estimation of nucleotide variability 60

4.3 The *ac-sc* region 61
4.3.1 Nucleotide variability 63
4.3.2 Nonrandom associations 63
4.3.3 Three-way associations 66
4.3.4 Haplotype diversity 66
4.3.5 Insertion-deletion variation 68
4.4 The *l*S12-*ry* region 68
4.4.1 Nucleotide variability 72
4.4.1.1 Distribution of site polymorphism 72
4.4.1.2 Heterozygosity 72
4.4.2 Nonrandom associations 73
4.5 Comparison of the two regions 78

5 DISCUSSION **80**

5.1 Early observations of disequilibrium 81
5.1.1 Morphological characters 81
5.1.2 Gene complexes: the *HLA* system 82
5.1.3 Inversions and allozyme loci 84
5.2 Models of multilocus evolution 85
5.3 Test of the theory: allozyme data 85
5.3.1 Plant populations 86
5.3.2 Microorganisms 86
5.3.3 Outbreeding animals: molluscs and vertebrates 87
5.3.4 *Drosophila* allozyme loci 87
5.4 Molecular genetics 88
5.4.1 Human gene complexes 89
5.4.2 *Drosophila* genes 90
5.5 Overview 91
5.6 Features of the results 92
5.6.1 Disequilibrium values 92
5.6.2 Heterozygosity estimates 92
5.7 Implications of the results 94
5.7.1 Distribution of disequilibrium 94
5.7.2 Evolution of the *ac-sc* complex 95
5.8 Limits of the analysis 96
5.8.1 Autocorrelation of data 96
5.8.2 Experimental technique 97
5.9 Suggestions for further work 98
5.10 Current perspectives 99
5.11 Conclusions 100

ACKNOWLEDGEMENTS **101**

REFERENCES **102**

APPENDIX **114**

I Units and Nomenclature **115**

II Solutions and media **116**

III Fly lines **120**

IV Summary of data **123**

CHAPTER 1

INTRODUCTION

1.1. Linkage disequilibrium

Some knowledge of the population-genetic behaviour of multilocus systems is essential to a full understanding of the evolutionary process, since no part of the genome can evolve in a way which is entirely independent of its genetic background. The arrangement of the genetic material into chromosomes ensures that the relationship between a large fraction of the genes is not only functional but also physical; linkage of genes on different chromosomes is broken down by independent assortment of the chromosomes at meiosis, while that of genes on the same chromosome is destroyed only by crossing-over between homologous chromosomes at a greater or lesser rate dependent upon the distance separating them. These two events together constitute recombination, the major promotor of genetic diversity in sexually-reproducing organisms.

The population genetics of individual loci was first addressed by Hardy (1908) and Weinberg (1908) who showed that in random-mating populations without selection gene frequencies remain unaltered through time. This rule when extended to cover multiple loci is profoundly affected by linkage (Jennings, 1917). When more than one locus is considered, even if each separately is in Hardy-Weinberg equilibrium, one cannot describe the genetic variability of a population solely in terms of the individual allele frequencies at each locus: superimposed upon this variation is the extent of non-random association between particular alleles at different loci. If some combinations of alleles are more or less common in a population of haploid gametes than their random expectations, the loci involved are said to be in *linkage disequilibrium* (Lewontin and Kojima, 1960). This quantity is normally defined as the gametic excess of 'coupling' over 'repulsion' genotypes; hence for two loci A and B each polymorphic for two alleles (A_1 and A_2; B_1 and B_2 respectively) the *coefficient of linkage disequilibrium* is given by:

$$D = f(A_1B_1)f(A_2B_2) - f(A_1B_2)f(A_2B_1)$$

The frequency of each of these four gametic classes will then differ from the

product of its individual allele frequencies by an amount equal to $\pm D$, thus:

$$f(A_1B_1) = p_1q_1 + D \qquad \text{where } p_1 = f(A_1)$$
$$f(A_2B_2) = p_2q_2 + D \qquad p_2 = f(A_2)$$
$$f(A_1B_2) = p_1q_2 - D \qquad q_1 = f(B_1)$$
$$f(A_2B_1) = p_2q_1 - D \qquad q_2 = f(B_2)$$

The parameter D may be thought of as the interaction deviation which together with p_1 and q_1 completes the three degrees of freedom of the gametic frequencies at the two loci (Lewontin, 1974). For more than two loci, higher-order interactions can exist: hence for three loci there are eight gametic frequencies, seven of which are independent; they can be fully described by three main effects (the allele frequencies at each locus), three first-order interactions (the D values for the three pairwise combinations of the loci) and a second-order interaction, the deviation of the gamete frequencies from random combination not accounted for by the pairwise D values (Bennett, 1954; Hill, 1974a).

Similarly, third-order disequilibrium additional to the pairwise and three-way values may be established across four or more loci. When linkage disequilibrium extends across many loci in this fashion, only a fraction of all multilocus arrays or 'haplotypes' will be represented in the gamete population.

Several other measures of disequilibrium have since been described (see Hedrick 1985, 1987; Hedrick *et al.*, 1978) which provide useful alternatives to D in certain contexts. Of these the most widely quoted are D', the ratio of D to its maximum value for the given allele frequencies (Lewontin, 1964), and the squared correlation coefficient r^2, equal to $D^2/p_1p_2q_1q_2$ (Hill and Robertson, 1968) both of which take preference over D by virtue of their reduced dependence on allele frequencies. The relative merits of these three measures are considered further in Chapter 4.

The persistence of disequilibrium is above all dependent upon the rate of recombination between the loci. In the absence of forces to maintain it, the disequilibrium will decay in an exponential fashion with succeeding generations, each generation of recombination bringing the adjacent variants closer to random (equilibrium) combination (Geiringer, 1944). After t generations of random mating the expected disequilibrium between any two

loci is the following fraction of its initial value:

$$D_t = (1-c)^t D_0$$

where c is the recombination fraction between the loci. From this expression the theoretical pattern of decay of D can be derived for any value of c, as shown in Fig. 1. Note that linkage disequilibrium can be present and persist for a short period even with independent assortment (c=0.5).

The primary force responsible for the origin of linkage disequilibrium is mutation, as it produces a continual input of unique variants to the population each of which is initially in maximum disequilibrium with all neighbouring polymorphic sites, before recombination gradually destroys the associations. Such disequilibria, although maximal for the allele frequencies (maximum D') are nevertheless small in their absolute value, as would be measured by D or r^2, at the outset. However, as the mutant frequency increases, whether by genetic drift or selection, the absolute disequilibrium with the most tightly linked flanking markers will be enhanced at a faster rate than its decay by recombination. In finite populations the random sampling of alleles is expected to create variable amounts of disequilibrium, even though the expected average D is zero (Hill and Robertson, 1968; Ohta and Kimura, 1969). Hence, although mutation alone would be an unimportant source of disequilibrium, significant associations will appear over time by the combined action of mutation and drift.

Much higher initial levels of absolute disequilibrium may be generated by migration between adjacent populations which differ substantially in allele frequencies; the magnitude of these will depend upon the proportion of migrants m_x and m_y from the two parent populations respectively and the difference in equivalent allele frequency between them at each locus, in the following way:

$$D = m_x m_y (p_x - p_y)(q_x - q_y)$$

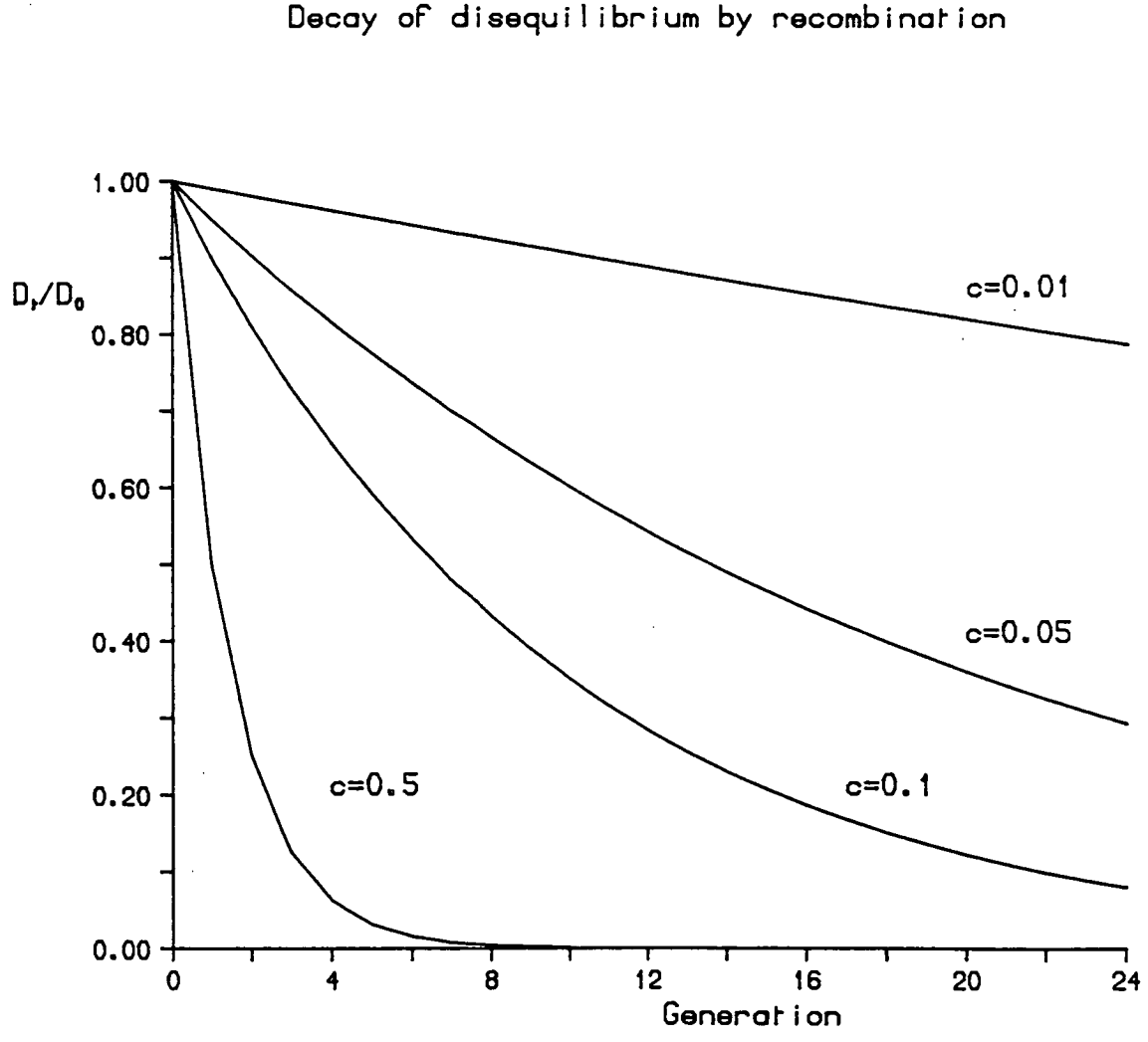
assuming D=0 in both the parent populations (Cavalli-Sforza and Bodmer, 1971, p.69). In the extreme, admixture in equal proportions of populations fixed for alternative alleles at two loci would create the maximum

Figure 1

Rates of decay of linkage disequilibrium from an initial value, D_0 for various values of the recombination fraction c between pairs of loci.

Opposite page 4

Fig. 1


Decay of disequilibrium by recombination
1.00
0.80
0.60
0.40
0.20
0.00
D_t/D_0
c=0.01
c=0.05
c=0.1
c=0.5
0
4
8
12
16
20
24
Generation

disequilibrium (D=0.25). Thus migration may be an important source of nonrandom associations in species with a subdivided population structure (Ohta, 1982).

Of most interest to many evolutionists, though, is the possible generation of linkage disequilibrium by natural selection in favour of certain multilocus genotypes. This could take the form either of simultaneous directional selection on neighbouring loci, the genetic load due to two deleterious alleles being much reduced if they are in strong coupling disequilibrium; or an adaptive interaction between loci in determining fitness. (I use 'interaction' rather than the popular 'epistasis' to describe this latter category of selection to avoid confusion with the classical meaning of epistasis, *i.e.* the masking of a phenotype at one locus by that of another).

The essence of fitness interaction is that two or more loci combine to produce a multigene phenotype whose fitness cannot be predicted by the component single-locus fitnesses, either on an additive or a multiplicative scale. The term 'supergene' was used by Darlington and Mather (1949) to describe 'any group of genes acting as a mechanical unit in particular allelomorphic combinations'. Numerous morphological traits have long been recognized as examples of such supergenes, many as a result of observations of linkage disequilibrium in data accumulated from field studies. The spatial distribution of morphs for shell colour and banding in *Cepaea* snails in relation to habitat type (Cain and Sheppard, 1954; Clarke, B., *et al.* 1968) is one famous instance, the multiple loci controlling this composite trait presumed to be held in disequilibrium by a common selection pressure, that of predation. At each locus separately, gene frequencies are thought to be kept intermediate by balancing selection as a consequence of habitat heterogeneity; in addition, certain combinations of colour and banding are considerably more common against particular substrates even when recording bias is minimized by the intensity of surveying. Hence loss of fitness could conceivably accrue not only from mutation at any one of the controlling loci, but more often by recombination between them. Other supergenes in this category have been inferred from the natural history of the primrose, *Primula* and the *Papilio* butterflies (see Discussion).

The influence of selection extends beyond its direct action on

genotypes. A neutral variant closely linked to a locus under (directional) selection will tend to change in frequency in parallel with it, the phenomenon of 'genetic hitchhiking' (Maynard Smith and Haigh, 1974). The hitchhiking effect can also be observed as an apparent heterosis of a neutral allele linked to an overdominant locus, where it has been called 'associative overdominance' (Frydenberg, 1963) or 'pseudoselection' (Thomson, 1977). Although at first sight hitchhiking appears to be a mechanism for generating disequilibrium, it is in fact dependent upon at least some initial disequilibrium between the loci. However, Thomson (1977) has shown that the simultaneous pseudoselection of two neutral loci in close linkage with an overdominant locus can create substantial disequilibrium between them even from an initial D=0, besides enhancing their respective disequilibria with the selected locus. Hence the effect of selection on linkage disequilibrium becomes impossible to disentangle from that of close linkage to selected loci.

Supergenes and other tightly linked gene complexes are obvious instances where linkage disequilibrium should be expected; but the outcomes of a series of early mathematical models incorporating linkage and heterotic selection suggested that if interactions for fitness are common between adjacent loci, then large-scale disequilibria involving most or all of a chromosome could be arrived at (Lewontin, 1964; Franklin and Lewontin, 1970; Slatkin, 1972). This could be attributed to two related phenomena: *cumulative effect,* in which a chain of interactions along a chromosome results in quite widely-displaced loci being held in disequilibrium; and *embedding effect,* whereby disequilibrium between tightly linked variants in a complex is magnified by their respective interactions with surrounding loci (Lewontin, 1974). In this way the summation of relatively weak individual fitness interactions across many loci could produce a large higher-order interaction. As we shall see, however, little evidence of such 'crystallization of the genome' can be found when natural populations are investigated experimentally (Langley, 1977; Clegg, 1978; Clegg *et al.,* 1980).

The experimental measurement of linkage disequilibrium is of course complicated by the difficulty of determining gametic frequencies in heterozygous diploid organisms, and procedures have been developed for estimating disequilibrium from phenotypic frequencies by maximum likelihood (Hill, 1974b). In *Drosophila,* however, the opportunity exists for the direct

identification of gametes by 'extracting' individual chromosomes with the aid of artificially-constructed, homologous 'balancer' chromosomes. These balancers typically feature multiple rearrangements, especially inversions, to suppress recombination over a substantial portion of the chromosome, and a series of visible genetic marker alleles to facilitate their ready identification (for a description of some common balancers see Lindsley and Grell, 1967).

1.2. Molecular variation

The measurement of genetic variability in populations has benefited considerably from the techniques of molecular biology, which allow both proteins and DNA fragments to be analysed for differences in their electrophoretic mobility (Lewontin and Hubby, 1966; Jeffreys, 1979), yielding an estimate of variation unbiased by its prior discovery. The views propounded by Kimura (1983) in the neutral theory of molecular evolution postulate that, at least at the molecular level, the vast majority of existing polymorphism may be maintained by stochastic processes (mutation and genetic drift) rather than, as had been widely assumed in the past, by balancing selections (Ford, 1965; Dobzhansky, 1970). The resultant controversy has given added impetus to the effort to accurately measure levels of naturally-occurring molecular variation, in order to test predictions arising from the neutral theory.

Variation in protein sequences can be detected as a difference in the electrophoretic mobility of a polypeptide chain, arising either from the truncation of the chain or from the substitution within it of a different amino-acid with consequent alteration of its overall electrical charge. Application of electrophoresis to natural population samples initially indicated that as many as one in three loci in an average species could be polymorphic in either of these two ways (Harris, 1966; Lewontin and Hubby, 1966) providing for the first time a relatively unbiased estimate of the degree of polymorphism in populations. The subsequent development of recombinant DNA techniques has allowed this variation to be related to its underlying cause, the primary pool of DNA sequence variability which stems from mutation.

Variation in nucleotide sequences can be of two forms: the simple substitution of one base-pair for another without alteration in the length of the DNA sequence, and variation caused by the insertion or deletion of segments of DNA which results in sequence-length changes of anything from

one to several thousand base-pairs. Natural populations, under the neutral theory, might be expected to support a higher fraction of the former class of variation as it is less likely to result in deleterious effects: single-site variants can arise in coding regions without affecting the amino-acid sequence of a protein if they occur in the third position of a codon ('silent' substitutions); moreover, single-base changes which do result in an amino-acid substitution often have little or no effect on the protein's activity and can therefore be accommodated without significant selective disadvantage. In contrast, most insertion-deletion events within a reading frame are likely to disrupt protein activity to a profound extent by generating a frameshift with consequent truncation of the reading frame, resulting in 'null' mutations with a high detriment to fitness. Hence, any sequence-length variation supported by natural populations would be expected to be found in non-coding regions of the genome.

Cloning and sequencing of the genomic DNA will reveal both classes of variation, but is a time-consuming and inefficient method of detecting variants over anything other than very short stretches of DNA in a very few individuals. The availability of restriction endonucleases from bacteria which cleave DNA at sites with a specific short sequence of bases provides a convenient way of observing single base substitutions occurring within these cleavage sequences, as restriction sites are either created or abolished by such changes and a different pattern of fragments is therefore produced by electrophoresis. Sequence-length variation over and above a certain size will also be detected as a change in restriction-fragment lengths, and unlike base substitutions a consistent change will appear when several enzymes recognizing different sequences are used. The various adaptations of the Southern transfer hybridization (Southern, 1975; Wahl *et al.*, 1979; Kreitman and Aguadé, 1986) enable any region of the genome to be investigated for which an appropriate homologous probe is available (see Methods).

A greater fraction of the existing genetic variation can be uncovered by this approach than by the electrophoresis of proteins, which cannot detect silent substitutions and variation in non-coding DNA sequences, nor can it readily distinguish between base substitutions and sequence-length changes. The similarity of many allozyme electromorphs may detract further from the sensitivity of protein-based surveys by pooling several distinct alleles into one

mobility class. This problem, besides adversely affecting estimates of polymorphism, could also conceal many instances of linkage disequilibrium between the pooled variants since positive and negative disequilibria within a class will be liable to cancel out (Zouros *et al.,* 1977; Weir and Cockerham, 1978). The extent of this effect will vary depending upon the size and complexity of the particular proteins being assayed.

Of late, therefore, DNA-level studies have become the preferred method of screening for genetic variation, whether by sequencing (Kreitman, 1983) or by restriction-enzyme mapping (Jeffreys, 1979; Langley *et al.,* 1982; Leigh Brown, 1983; Aquadro *et al.,* 1986; Cross and Birley, 1986; Kreitman and Aguadé, 1986). The majority of such studies have concentrated on *Drosophila melanogaster* owing to the advanced genetic and molecular information available for this species.

1.3. Aim of the survey

The experimental study of linkage disequilibrium benefits from the ability of restriction enzymes to map the genome to a much finer scale than would be possible using gene product-dependent characters. Variation involving the very closest genetic loci can thus be resolved, where disequilibrium should be most evident; any change in disequilibrium over the transition to more distant loci can then be evaluated. The precise distances separating the variants will be known if loci that have been previously mapped with restriction enzymes are studied.

The objective of this survey is to quantify the relationship of linkage disequilibrium to the physical and genetic separation of variants along a chromosome. To achieve this, a comparison was initiated of DNA polymorphism at two well-characterized loci, the *rosy-Ace* and *achaete-scute* regions from a natural population sample of the fruit fly, *Drosophila melanogaster* Meigen. Both of these include numerous genetic and cloned markers and are also known to be subject to contrasting rates of recombination. The comparison of linkage disequilibrium between the regions will therefore be of considerable relevance in determining the regression of disequilibrium on recombination fraction for loci in natural populations, which in turn impinges upon the effort to identify natural selection at the multilocus level.

The strongest evidence for multilocus selection would be the discovery of statistically significant linkage disequilibria, other than those expected from tight linkage, which are consistent between populations (Lewontin, 1974). To recognize such associations requires an appreciation of the natural sampling distribution of linkage disequilibrium from a wide range of different loci and populations, since the current methods for establishing the statistical significance of nonrandom associations may have limited validity when very tightly linked loci are considered. Specifically, when $4Nc<10$ where N is the population size, the distribution of disequilibrium departs from that of chi-square (Golding, 1984). The comparison of disequilibrium data between loci and populations is thus necessary to establish independent assessment of its distribution and thereby circumvent the need for undue reliance on χ^2 values, while simultaneously yielding a result of more general validity across the genome.

Large sample sizes are also a key feature of this survey: as elaborated by Brown (1975) the sample sizes required to establish statistical significance of *D* rise dramatically the greater the discrepancy between allele frequencies at the two loci, the further they depart from the intermedium (0.5), and the lower the parametric (true) disequilibrium. For p=q=0.5, a sample size of around 30 will suffice to reject a null hypothesis of D=0 with 90% confidence for all values of *D'* greater than 0.5; but if *p* and/or *q* are around 0.2 or less, the number of random gametes (chromosomes) required to be 90% sure of rejecting D=0 can range from 80 to several thousand for moderate to low values of *D'* (see table II in Brown, 1975). As variants at intermediate frequencies are rare, and the establishment of many isochromosomal lines laborious, sample size has therefore been a severe limiting factor in many previous experimental studies of linkage disequilibrium.

1.4. Genetic natural history of the regions surveyed

The *yellow-achaete-scute* complex maps to the telomeric (1B1-1B4/5) region of the X chromosome (Carramolino *et al.*, 1982; Campuzano *et al.*, 1985; Biessmann, 1985; Chia *et al.*, 1986). The *y* gene controls cuticular melanin expression, while the *ac-sc* portion consists of a series of genes affecting the number and distribution of the micro- and macrochaetae, and also necessary for the full development of internal elements of both the

peripheral and the central nervous systems (Jiménez and Campos-Ortega 1979, 1987; Ghysen and Dambly-Chaudière, 1989). Although crossing-over was demonstrated for the first time between *ac* and *sc* (Dubinin *et al.,* 1937) it is now known that the rate of crossing-over is very low relative to other X-linked *Drosophila* loci: the quoted separation of the two genes by 6.6×10^{-3} map units represents a mapping function of 1.2×10^{-4} map units kb^{-1}, compared to 2×10^{-3} m.u. kb^{-1} at *white* (Langley and Aquadro, 1987) and 5×10^{-3} m.u. kb^{-1} at *Notch* (Kidd *et al.,* 1983). Even this low figure must be considered a maximum since stocks bearing autosomal inversions designed to enhance crossing-over on the X chromosome were used. More recent genetic analyses, by mapping the breakpoints of deletions, have led to the subdivision of the *ac-sc* complex into the four regions *achaete, scute-*α, *lethal of scute* (*l'sc*) and *scute-*β (García-Bellido, 1979); the alignment of these genetic loci and the positions of transcripts are shown in Fig. 2.

The other genomic region chosen for study was the *rosy-Ace* region of the 87D6-87E5 segment of the third-chromosome (Hall *et al.,* 1983; Spierer *et al.,* 1983; Bender *et al.,* 1983) which includes *rosy,* the gene for xanthine dehydrogenase whose many documented mutations result in deficiency of red pigment in the eye, with consequent red-brown eye colour phenotypes; the flanking lethal *l*S12; *piccolo,* mutants for which have short, fine bristles and tergites etched at the margins; and *Ace,* formerly *l*26, the gene for acetylcholinesterase (Fig. 3). The first three of these have been localized to within 20 kb of one another, but *pic* and *Ace* are separated by over 150 kb of DNA including one exceptionally large gene, *B*16-1, extending over 50 kb (Gausz *et al.,* 1986). Intensive fine-scale recombination analysis of *rosy* has brought the number of known complementation groups close to saturation and indicates a mapping function of 1.1×10^{-3} map units kb^{-1} (Schalet *et al.,* 1964; Chovnick *et al.,* 1976), at least an order of magnitude higher than at *ac-sc* (see above). In addition, the complete sequence of the *ry* gene has been determined (Keith *et al.,* 1987) which permits a fine-scale molecular investigation of this region.

Figure 2

Location of genetic units within the *yellow-achaete-scute* complex of the X chromosome in *Drosophila melanogaster*. Transcripts are shown as open boxes, with arrows indicating the direction of transcription where this is known. Map ordinates are those of Campuzano *et al.*, 1985.

Opposite page 12

Fig. 2

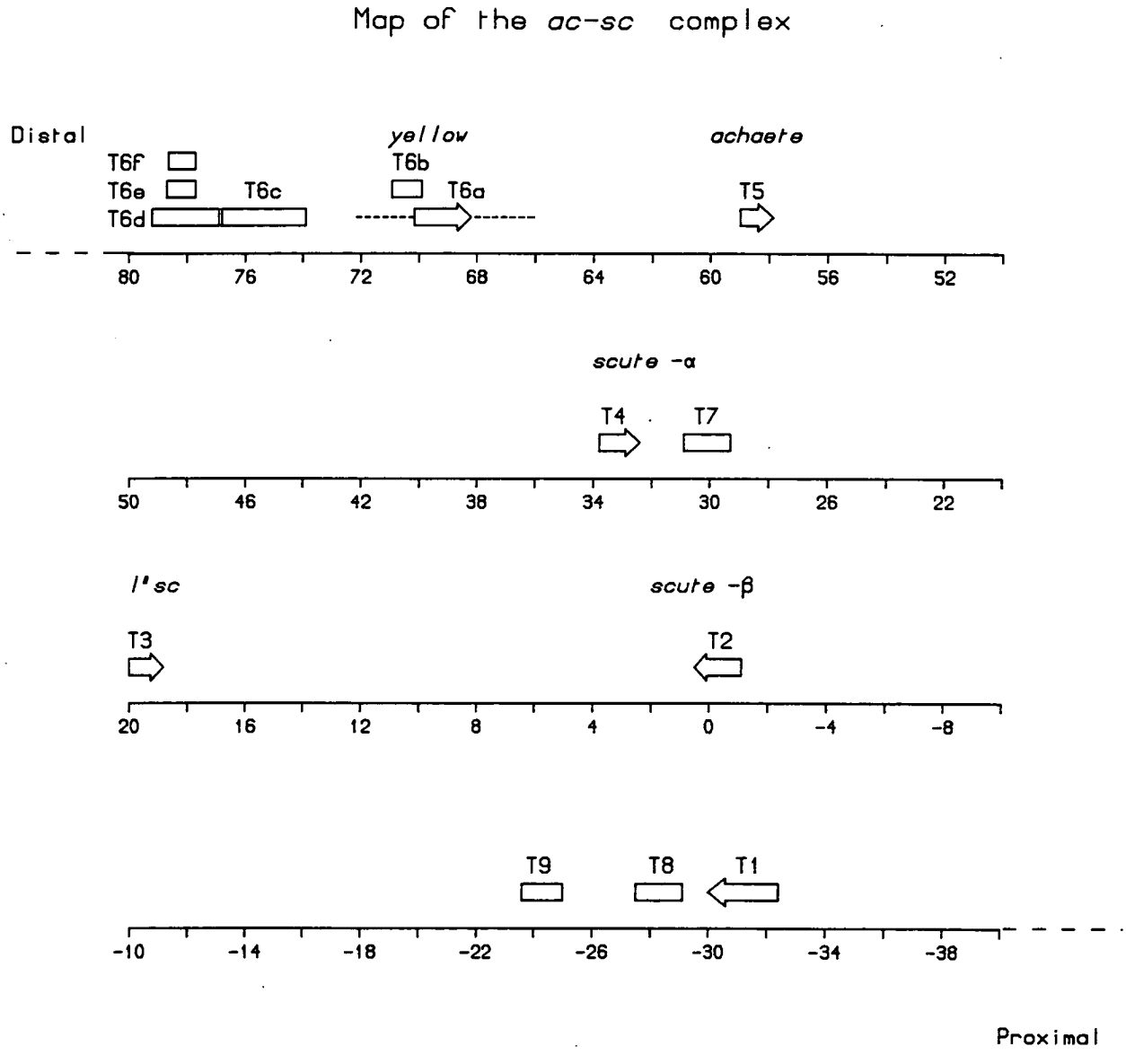

Map of the ac-sc complex
Distal
yellow
achaete
T6f
T6e
T6d
T6c
T6b
T6a
T5
80
76
72
68
64
60
56
52
scute -α
T4
T7
50
46
42
38
34
30
26
22
l'sc
scute -β
T3
T2
20
16
12
8
4
0
-4
-8
T9
T8
T1
-10
-14
-18
-22
-26
-30
-34
-38
Proximal

Figure 3

Positions of major genes and transcripts in the *rosy–Ace* region of chromosome III in *Drosophila melanogaster* (adapted from Hall *et al.*, 1983; Gausz *et al.*, 1986). The deficiency (3R)*kar*^{SZ11} has its distal breakpoint within the limits shown and extends proximally beyond -200 kb, thereby deleting the entire region. Map ordinates are those of Bender *et al.* (1983).

Opposite page 13

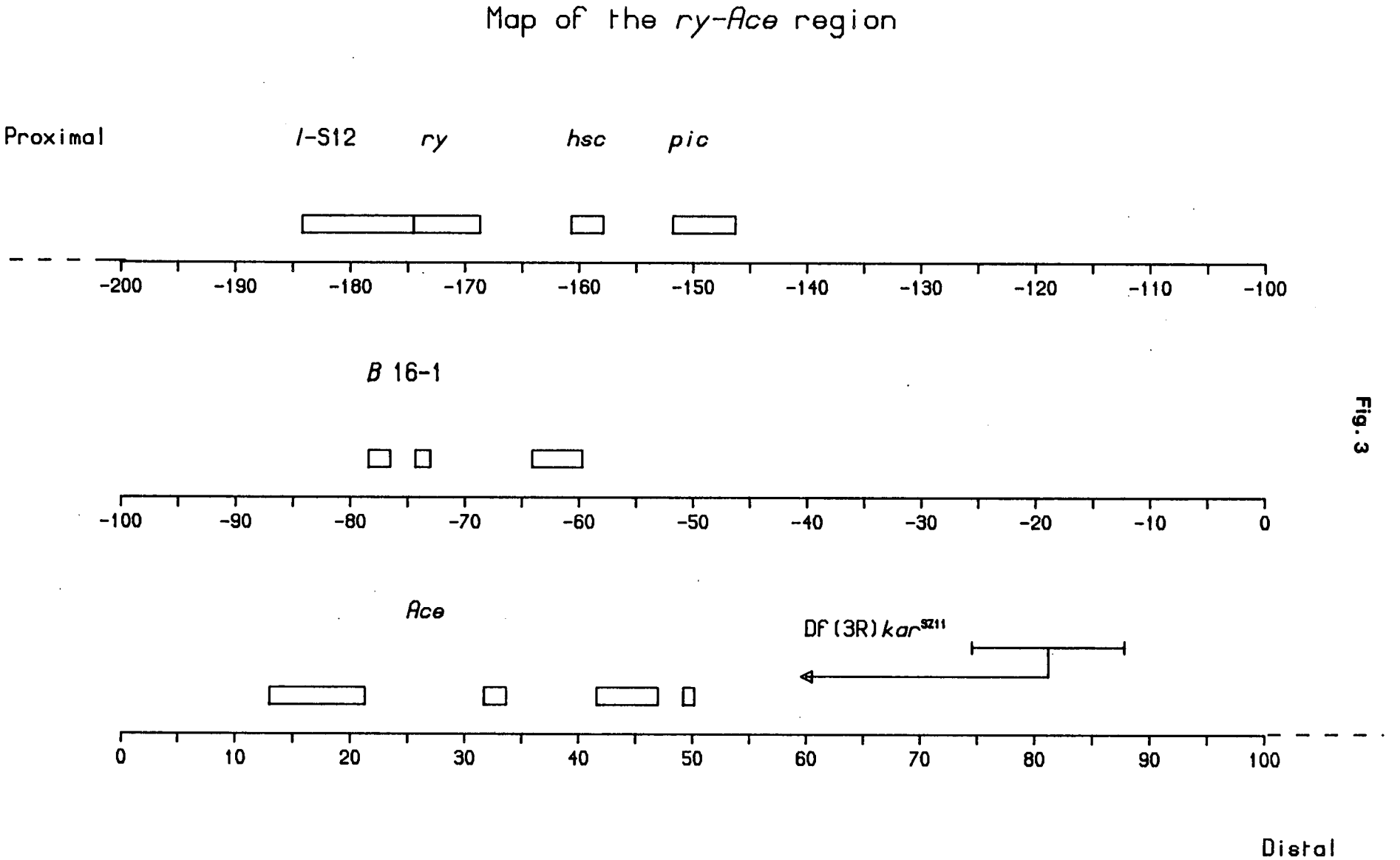

Map of the ry-Ace region
Proximal
l-S12
ry
hsc
pic
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
B 16-1
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
0
Ace
Df(3R)kar^SZ11
0
10
20
30
40
50
60
70
80
90
100
Distal

Fig. 3

1.5. Strategy for surveying the regions

A series of overlapping DNA sequences from each of the two regions was acquired as recombinant clones in bacteriophage vectors (Carramolino *et al.,* 1982; Campuzano *et al.,* 1985; Spierer *et al.,* 1983). These were used to derive appropriate short fragments (usually 2–4 kb in length) for use as homologous probes to genomic DNA samples from independent lines of flies. Each line had been made either homozygous or hemizygous for either the third- or the X chromosome. Genomic DNA was digested with an array of restriction enzymes and the allelic state of each fragment in the sample determined by standard methods of gel electrophoresis and Southern transfer.

The choice of enzymes has an important influence on the sensitivity of mapping: use of the half-dozen or so restriction enzymes recognizing tetranucleotide sequences, which therefore cleave on average 16 times more frequently than their counterparts recognizing hexanucleotide sequences, would screen more sites and in the process detect more polymorphisms than an equivalent number of six-cutters. The common four-cutter enzymes partition into two distinct categories: those which recognize sites composed of one each of the bases A, G, C and T, and those which recognize sites consisting of G and C only. Taking into account the base composition of *Drosophila* DNA (43% G+C), the overall frequency of cleavage sites for these enzymes (four in each class) may be estimated at 1/300 base-pairs = 3.3 sites kb^{-1}, hence 4 x 3.3 $\approx$13 nucleotides would be screened per kilobase probed, or over 200 nucleotides per 2 kb probe by the set of eight enzymes.

If the proportion of polymorphic nucleotides at the loci surveyed here was around 0.015, a value estimated for the *Adh* locus (Langley *et al.,* 1982) then these eight enzymes should between them detect an average of 200 x 0.015 = 3 polymorphic sites with each 2 kb probe used. In practice, however this is unlikely since many of these will involve fragments too small to be observed. Further, precise map positions are not commonly available for four-cutter sites unless the DNA sequence is known; for these reasons it remains desirable to use a combination of six- and four-cutter mapping to resolve variation at both fine and coarse levels.

The restriction-site data were used to estimate the average single-nucleotide heterozygosities of the two regions, and these results

compared with previous estimates from the same North Carolina population (Beech and Leigh Brown, 1989; Aquadro *et al.,* 1988). Linkage disequilibrium was calculated by standard means for every possible pairwise combination of the variants observed, and the regression of disequilibrium on distance constructed and compared between the two regions. If we are dealing with an ideal Hardy–Weinberg population (*i.e.* large and panmictic, without selection) then pairwise linkage disequilibrium should decline progressively with increasing recombination interval between variants; consequently, higher disequilibrium should be expected across greater distances in the *ac–sc* region where the rôle of recombination is curtailed.

Once all possible pairwise disequilibria had been fully evaluated, the data were re–examined for the presence of higher–order associations which, if detected, would form the basis for further speculations concerning the evolutionary history of the regions.

CHAPTER 2

METHODS

The compositions of solutions and media are detailed in Appendix II.

2.1. Bacterial stocks

Strains of *Escherichia coli* were maintained on plates of suitable agar medium (usually L-agar) in Petri dishes at 4 °C for purposes of short-term storage. Fresh single colonies were obtained on such plates by streaking bacteria either from a single colony on an existing plate or from a stock culture and growing at 37 °C overnight. For immediate use, liquid cultures were prepared by inoculating 10 cm^3 of suitable broth with a single fresh colony and incubating at 37 °C overnight with shaking. Larger-scale cultures could then be set up if necessary by inoculating up to 500 cm^3 of broth with the cell suspension at a density of around 1% by volume. Long-term storage of bacterial stocks was accomplished by withdrawing 0.5 cm^3 of a 10 cm^3 overnight liquid culture and mixing thoroughly with an equal volume of sterile propane-1,2,3-triol (glycerol). Stored at -20 °C such stocks retain viability for several years.

E. coli strain Q358 (Karn *et al.*, 1980) was used as a host for infection with bacteriophage λ. Plating cell stocks were prepared by inoculating 10 cm^3 of T-broth containing 2% maltose with a fresh single colony, incubating with shaking at 37 °C overnight and sub-culturing at 1% by volume in a further 10 cm^3 of T-broth + maltose for 6 h. The cells were then centrifuged at 1 500 *g* for 10 min and resuspended in 5 cm^3 of 10 mM $MgCl_2$, in which condition they remain viable for about a week. The maltose supplement improves adsorption of the phage to the bacterium as the receptor for λ adsorption is encoded by the maltose operon (Thirion and Hofnung, 1972) and Mg^{2+} ions must be present throughout to maintain the integrity of the phage particles (Maniatis *et al.*, 1982).

2.2. Bacteriophage stocks

Bacteriophage λ is normally stored for up to several years as the supernatant from a liquid lysate of *E. coli*. The cell debris is spun down at 1 500 *g* for 10 min and the supernatant transferred to a sterile glass vial to

which a drop of trichloromethane (chloroform) is then added to prevent bacterial growth. Single plaques may be obtained from such lysates by infecting 0.1 cm^3 of Q358 plating cells with an appropriate dilution of the lysate and allowing the phage to adsorb at 37 °C for 20 min before mixing thoroughly with 3 cm^3 of L-top agar at 50 °C and pouring immediately over the surface of an L-agar plate, which is then incubated at 37 °C overnight. Liquid lysates with a high titre of plaque-forming units (up to 10^{12} p.f.u. cm^{-3}) can be obtained at any time from these plates by removing a single plaque with a sterile Pasteur pipette and inoculating the plug of agar containing the plaque together with 0.1 cm^3 of plating cell suspension into 10 cm^3 of L-broth, shaking vigorously for 8 h until lysis is complete and removing the cell debris as described above.

2.3. *In vitro* packaging of bacteriophage DNA

Precipitates of phage DNA in ethanol were spun down in a microcentrifuge and resuspended in 50 mm^3 TE. Assuming a packaging efficiency of 10^5 plaques μg^{-1} DNA, around 50 ng of each sample was added to 6 mm^3 of sonicated extract and 10 mm^3 of freeze-thaw lysate prepared as described in Scalenghe *et al.* (1981). The reaction was left at room temperature for 1 h before being halted by the addition of 100 mm^3 phage suspension buffer (PSB) and stored at 4 °C. An appropriate quantity of the packaged sample was then used to produce single plaques on a Q358 host as described above.

2.4. Large-scale preparation of bacteriophage DNA

Large-scale preparations of λ DNA, or 'maxipreps' were initiated from liquid lysates with a titre of around 10^9 p.f.u. cm^{-3}. Overnight liquid cultures of host Q358 cells were used to inoculate 400 cm^3 of L-broth, minus glucose but supplemented with 10 mM Mg^{2+}, to an optical density at 600 nm ($O.D._{600}$) of 0.1. The cells were shaken at 37 °C until the $O.D._{600}$ reached 0.4, when about 10^9 p.f.u. of phage was added and the cultures shaken vigorously for a further 4-6 h until lysis was achieved. Omitting glucose from the medium elevates the expression of the maltose operon, thereby enhancing phage adsorption.

The lysed culture was shaken for a further 15 min with the addition of

0.1 cm^3 trichloromethane, 10 µg cm^{-3} DNAse and 10 µg cm^{-3} RNAse (Sigma) to ensure complete disintegration of the cells. The lysate was made 1 M with NaCl and centrifuged at 16 000 *g* for 10 min to separate cell debris; to the supernatant was added 7% solid crushed *poly*ethane-1,2-diol, RMM 8 000 (polyethylene glycol, PEG; Sigma) before leaving at 4 °C for at least an hour or overnight. Precipitated phage particles were spun down at 10 000 *g* and the pellet resuspended gently in 5 cm^3 of TMN. Following two extractions with an equal volume of trichloromethane/3-methylbutanol, spinning at 12 000 *g* for 10 min each time, the phage suspension was made 41.5% CsCl in TMN and centrifuged to equilibrium at 80 000 *g* overnight.

The bluish band of phage particles was collected in a siliconized Pasteur pipette and dialysed against 10 mM Tris pH 8, 5 mM NaCl, 1 mM ethane-1,2-diamine-N,N,N',N'-tetraethanoic acid (EDTA) to remove CsCl. Phage particles are sensitive to EDTA and in the absence of Mg^{2+} tend to fall apart, so the DNA was readily purified by three equal-volume extractions, first with buffered phenol, then with phenol-trichloromethane and finally with trichloromethane/3-methylbutanol, and precipitated by adding 0.1 volume ammonium ethanoate and 2 volumes ethanol. Yield was typically 100 µg.

2.5. Marked *Drosophila* stocks

Stocks of *Drosophila melanogaster* incorporating a suitable balancer chromosome were made use of in the extraction of wild third- and X chromosomes. The balancer chromosome serves simultaneously as a non-recombinable homologue to the wild-derived chromosome, and as a marker for the balanced line, since it contains several mutant alleles. The stocks used were *TM6B, Tb e*/π_2 third-chromosomes and *C(1)DX, y f*/*sn*^{w} X chromosomes, both donated by Dr W.R. Engels. Both are strong P strains as they have essentially a π_2 background (Engels 1979, 1985) ensuring that there is no induction of hybrid dysgenesis upon crossing to wild strains which are also P cytotype.

An additional stock of genotype *TM6B, Tb e*/*mwh e* was constructed by Dr A.J. Leigh Brown by crossing virgin *TM6B*/π_2 females to male *mwh* (*multiple wing hairs*), *e* homozygotes. Progeny of the required genotype, recognizable by their *ebony* phenotype, were mated together to set up the final stock which was maintained for 5-10 generations before being used in

the third-chromosome extraction procedure, by which time it should have converted to P cytotype (Kidwell *et al.*, 1981); this was confirmed by the procedure described in 2.8 (Beech, 1987).

2.6. Chromosome extraction procedure

Fig. 4 shows the crossing scheme followed to establish the balanced third-chromosome lines; lethal progeny classes are omitted for simplicity. All crosses were carried out by J.E. Moss. Single wild-caught inseminated females, captured at Farmer's Market in Raleigh, NC, USA over a period of one week in 1984 by Dr A.E. Shrimpton, were used to found a series of isofemale lines from which single males were isolated and each crossed to several virgin females from the *TM6B*/π_2 stock. Single *Tubby* male progeny each possessing one wild-derived third-chromosome and the *TM6B* balancer third-chromosome were then crossed to several virgin *TM6B*/*mwh e* females. Of the four classes of progeny generated by this cross, the *TM6B* homozygotes are not viable, *TM6B*/*mwh e* are *ebony* in phenotype and flies heterozygous for the wild chromosome and *mwh e* are wild-type. The remaining class, with the identical wild chromosome balanced against *TM6B*, display a *Tubby* phenotype. The extracted third-chromosome line was therefore initiated by crossing together one pair of *Tubby* non-ebony flies; 72 independent lines were established in this fashion.

The X chromosome extraction procedure used isofemale lines from the same North Carolina population, this time crossing single first-generation males to *C(1)DX, y f* virgin females. Progeny from this cross were mated together to set up the extracted X chromosome line (Fig. 5); 44 lines were thus obtained, 27 of which were constructed by Robin Beech (Beech, 1987). The sex chromosomes in these attached-X strains exhibit a modified inheritance pattern whereby the female inherits an attached-X from its mother and a Y from its father, while the male inherits a normal X chromosome from its father and a Y from its mother. Hence the wild-derived X chromosome in this case is present only in males (unless the attached-X chromosome breaks) and therefore undergoes no recombination.

Figure 4

Crossing scheme used to establish extracted third-chromosome lines of flies (III_1/ *TM6B*) from which homozygotes for the third-chromosome and (right of dotted line) hemizygotes for lethal-bearing third-chromosomes were then derived (see 2.5–2.7 in text).

Opposite page 20

Fig. 4

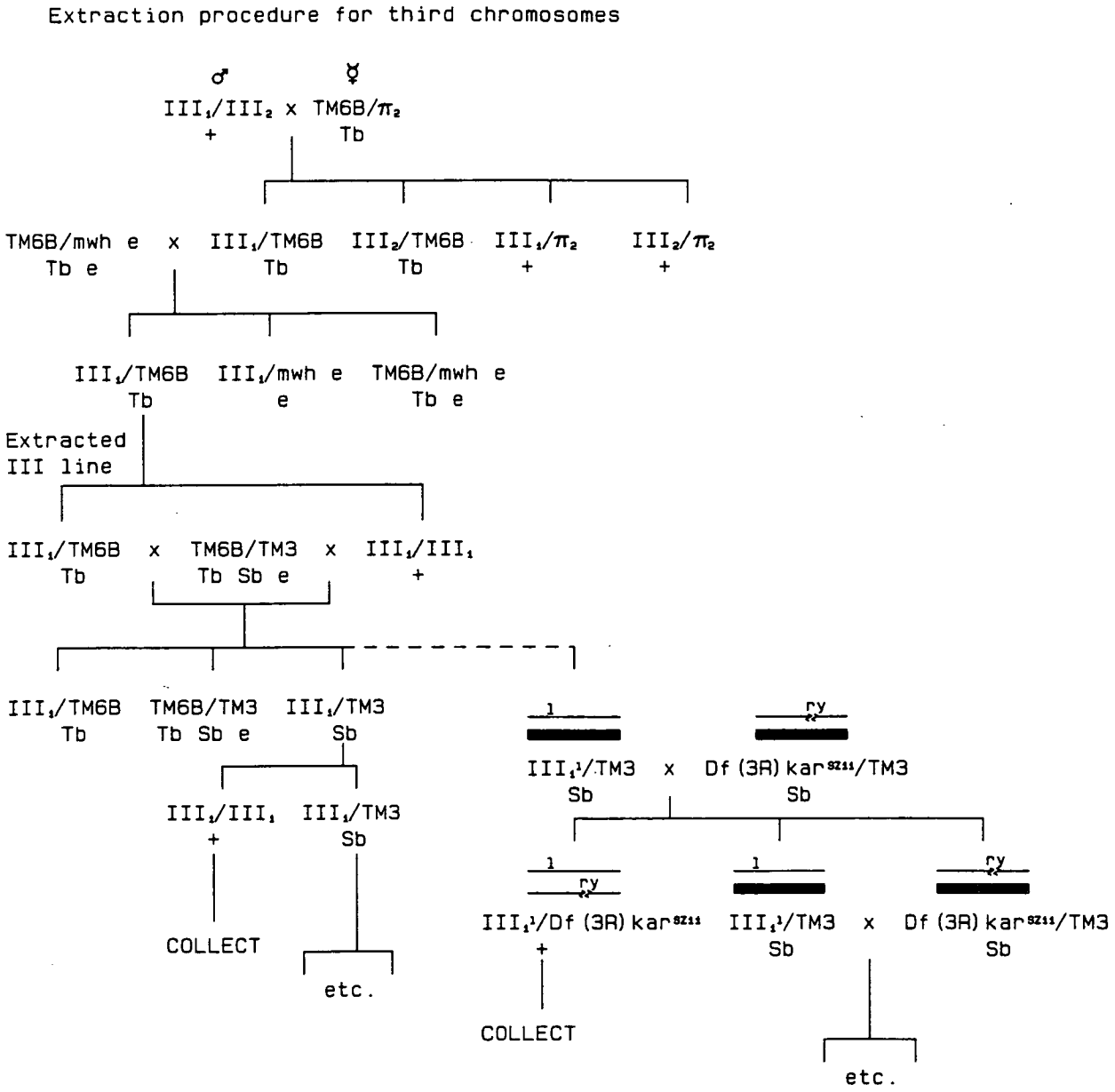

Extraction procedure for third chromosomes
♂
III1/III2 x
+
♀
TM6B/π2
Tb
TM6B/mwh e
Tb e
x
III1/TM6B
Tb
III2/TM6B
Tb
III1/π2
+
III2/π2
+
III1/TM6B
Tb
III1/mwh e
e
TM6B/mwh e
Tb e
Extracted
III line
III1/TM6B
Tb
x
TM6B/TM3
Tb Sb e
x
III1/III1
+
III1/TM6B
Tb
TM6B/TM3
Tb Sb e
III1/TM3
Sb
III1/III1
+
III1/TM3
Sb
COLLECT
etc.
1
ry
III1[1]/TM3
Sb
x
Df (3R) kar[sz11]/TM3
Sb
1
ry
III1[1]/Df (3R) kar[sz11]
+
1
III1[1]/TM3
Sb
ry
x
Df (3R) kar[sz11]/TM3
Sb
COLLECT
etc.

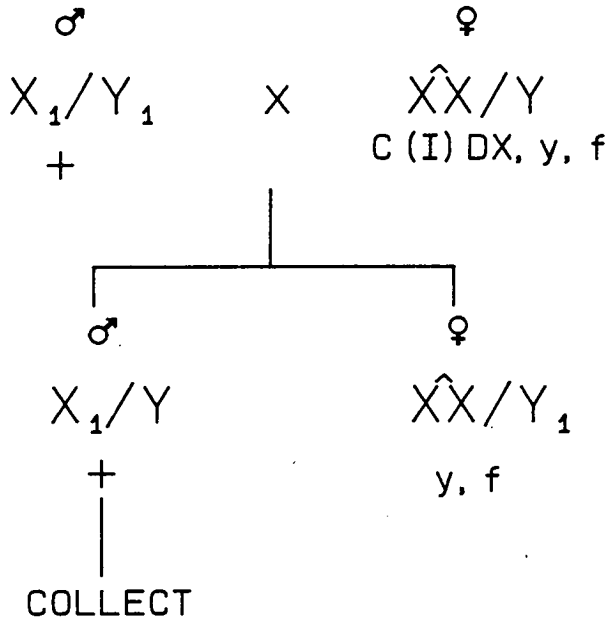
Figure 5

A simple crossing procedure to set up a stock of extracted X chromosome lines of flies. The attached-X in the female ensures that the wild-derived X_1 chromosome is transmitted through males only, without recombination (see 2.5-2.7 in text).

Opposite page 21

Fig.5

Extraction procedure for X chromosomes



2.7. Purification of loci

For purposes of Southern transfer hybridization it simplifies matters if the loci of interest are not heterozygous, since the balancer chromosome will also contain sequences homologous to the probes. For the X chromosome, this can be readily circumvented by collecting males only. In the case of the third-chromosome stocks, wild-type homozygotes cannot be identified with certainty owing to the difficulty in scoring the *Tubby* phenotype; accordingly an alternative balancer chromosome, *TM3,* incorporating the dominant marker *Stubble* (*Sb*) was utilized.[1] A P-cytotype *TM6B, Tb e / TM3, Sb Ser e* stock was supplied by Dr T.F.C. Mackay. In the crossing scheme outlined in Fig. 4, single males from the extracted III/ *TM6B* line (either *++* or *Tb+* in genotype) were mated to several virgin *TM6B / TM3 ebony* females and *Stubble* non-ebony progeny mated together to set up a balanced extracted III/ *TM3* line. Homozygotes for the wild third-chromosome were then recognized and harvested as wild-type flies in subsequent generations.

In almost half of the available lines, a recessive third-chromosome lethal or sublethal allele was indicated by the absence of wild-type flies in the F_2; it was therefore necessary to mate such lines to a stock with a third-chromosome deficient in the region of interest in order to purify the relevant wild sequences. The stock *Df(3R)kar*^{SZ11} / *TM3,* in which both *rosy* and *Ace* are deleted, was kindly provided by Dr J. Gausz (Gausz *et al.,* 1979); this was converted to P cytotype by crossing to (P) *TM6B / TM3* for two generations and subsequently mass-mating all progeny together for six generations, after which *Stubble* non-ebony progeny were mated with one another to reconstitute the *kar*^{SZ11} / *TM3* stock.

Mating of single extracted III/ *TM3* males to *kar*^{SZ11} / *TM3* females gives rise to only two viable phenotypes in the progeny: *Stubble*, which may be either III/ *TM3* or *kar*^{SZ11} / *TM3* in genotype, and wild-type flies of genotype III/ *kar*^{SZ11} in which the *ry-Ace* region is hemizygous with the deficiency. Wild-type flies were harvested *en masse* and stored at -70 °C prior to DNA

---

[1] V. Tinderholt, 1960. Drosophila Information Service 34, 53.

preparation.

2.8. Cytotype testing

The design of crossing schemes involving *Drosophila* is constrained by the P-M system of hybrid dysgenesis characteristic of this genus. Wild populations of flies in many of the world's zoogeographical regions, including the Nearctic, possess high copy numbers of the active transposable genetic element designated P (Bingham *et al.,* 1982). By contrast, many laboratory stocks were founded from wild strains lacking P elements; maintained free from contamination by flies with active P elements these strains, without any sequences homologous to the P, are said to display the M cytotype. The P-M nomenclature derives from the phenomenon observed when P and M strains are crossed reciprocally: when male (Paternal) flies from a P strain are crossed to female (Maternal) flies of an M strain, a diverse syndrome of effects is observed in the F_2 generation including greatly enhanced rates of mutation and chromosomal aberration, male crossing-over and temperature-sensitive sterility (Kidwell *et al.,* 1977) thought to be indicative of the mobilization of the active P elements in the germ line of the F_1 after their having been quiescent by virtue of being at their maximum copy number in the parent P strain (Bingham *et al.,* 1982; Rubin *et al.,* 1982). This syndrome, much less pronounced in the reciprocal cross between an M-cytotype male and a P-cytotype female, is obviously highly undesirable in any cross intended to preserve the integrity of the genome or even, as in this case, small segments of individual chromosomes. As the wild population employed in the survey will certainly be of P cytotype, hybrid dysgenesis can only be averted by saturating the balancer stock with P elements beforehand, as described above (2.7).

The cytotype of a particular *Drosophila* stock may be conveniently ascertained by mating to known P and M strains carrying the unstable mutation sn^w, and observing the F_2 progeny for signs of an enhanced mutation rate to sn^+ or sn^e characteristic of hybrid dysgenesis (Engels, 1979). Males of the stock to be tested were mated to about 100 virgin females of the M-cytotype *y* sn^w *bw st* stock, larvae were reared at 18 °C and male progeny mated to virgin *C(1)DX* females. About 400 F_2 males were scored for their *singed-bristle* phenotype; if the unknown stock is P then around 15%

combined sn^+ and sn^e would be expected.

The converse test should also be performed by mating 30–50 virgin females of the unknown stock to (P)*C(1)DX* males, rearing the larvae at 18 °C and mass-mating the adult F_1 progeny. In this case wild-type revertant males are indistinguishable from males carrying the grandmaternal X chromosome, so the frequency of sn^e revertants alone was scored. This would be expected to be negligible (<1%) if the tested stock is P.

2.9. Preparation of genomic DNA

Samples of around 500 flies from each line were thawed on ice before being ground up with several passes of the pestle in a manual glass homogenizer (Kontes Glass Co.) in 4 cm^3 of 10 mM Tris pH 7.5, 100 mM NaCl, 10 mM EDTA, 0.15mM N-[3-aminopropyl]-butane-1,4-diamine (spermidine) and 0.15mM N,N'-*bis*[3-aminopropyl]-butane-1,4-diamine (spermine). An equal volume of 200 mM Tris pH 9 containing 30 mM EDTA, 2% sodium dodecyl sulphate (SDS) and 0.2 mg cm^{-3} pronase-E (Sigma) was added and mixed gently with the homogenate before incubating at 37 °C for 1 h, by which time the nuclei should have lysed and most of the proteins been digested by the protease. The resultant DNA solution was separated from contaminating protein material by equal-volume extraction twice with buffered phenol, once with phenol-trichloromethane and once with trichloromethane/ 3-methylbutanol, made 0.7 M with ammonium ethanoate and the DNA precipitated by adding 2 volumes ethanol.

The precipitate was centrifuged at 4 000 *g* for 30 min to produce a pellet which, after removal of the supernatant, was allowed to dissolve in 2 cm^3 TE at room temperature overnight then reprecipitated as before. The DNA was then removed by spooling onto a long glass Pasteur pipette, the end of which had been heat-sealed to prevent capillary uptake of ethanol. After rinsing in 70% ethanol by volume to remove any traces of salt and drying thoroughly under vacuum, the DNA was finally redissolved in 1 mm^3 TE per fly to give a concentration of approximately 200 ng mm^{-3}.

2.10. Digestion of DNA with restriction endonuclease

DNA (usually 0.2–1.0 µg) was routinely digested with 10–20 units of restriction enzyme in one of the five buffers recommended by the manufacturers (Boehringer-Mannheim) at the appropriate temperature for over 4 h (genomic DNA) or for 2–3 h in the case of cloned DNA, by which time digestion should be complete. For most purposes, fragments of digested DNA were separated by horizontal agarose gel electrophoresis (McDonell *et al.*, 1977) in TBE buffer alongside marker DNA samples with fragments of standard size (usually bacteriophage λ digested with *Pst*I) and visualized by illumination with short-wave (260 nm) UV once the gel had been equilibrated with 1 µg cm^{-3} 2,7-diamino-10-ethyl-9-phenylphenanthridium bromide (ethidium bromide).

2.11. Agarose gel electrophoresis and transfer

Two adaptations of the method of Southern (1975) were utilized in the transfer of DNA fragments to nylon. To resolve fragments around 1–2 kb in length, digested samples were electrophoresed in high density gels of 1.8% agarose in TBE buffer which were run at 120 V for an appropriate time, soaked in 0.5 M NaOH, 1.5 M NaCl to denature the DNA then neutralized in 1 M ammonium ethanoate, 0.02 M NaOH. The gel was then supported on a wick of Whatman No.17 filter paper saturated with ammonium ethanoate/NaOH in which its ends were immersed. A sheet of 'Hybond' nylon membrane (Amersham International) was laid on top of the gel and a sheet of Whatman No.3 filter paper cut to the size of the gel, soaked in ammonium ethanoate/NaOH and laid on top of the 'Hybond'. Two sheets of dry filter paper and a 5 cm thickness of paper towels were added, taking care to prevent any contact between filter papers above and below the gel. The weight of a glass plate was then laid on top to enhance the upward movement of solution through the gel and papers.

Fragments 3 kb or greater in length were separated using gel concentrations between 0.3 and 1.2% agarose in TBE, run at 120 V for an appropriate time and treated with 0.2 M HCl for 20 min to partially cleave the larger fragments and facilitate their easier transfer. The DNA was denatured as above and neutralized in 1 M Tris pH 7.5, 1.5 M NaCl for 1 h before supporting the gel on a sheet of No.17 filter paper soaked in 20 x SSC and overlaying in

order 'Hybond', one sheet of No.3 filter paper soaked in 2 x SSC, two dry filter papers, paper towels, a glass plate and finally a light weight.

When transfer was virtually complete (after 1 h or overnight), membranes were rinsed in 2 x SSC to remove any traces of agarose, blotted dry and baked at 80 °C for 2 h to bind the DNA firmly to the nylon.

2.12. *poly*Propenamide gel electrophoresis and transfer

Fragments smaller than 1 kb in length were separated in vertical gels of 7% *poly*propenamide (polyacrylamide) of the type used in DNA sequencing (Church and Gilbert, 1984). Samples were first precipitated in ethanol and resuspended in 4 mm^3 of loading buffer comprising 96% methanamide, 20 mM EDTA, 0.25% Xylene Cyanol dye and 0.25% Bromophenol Blue dye (BDH), then denatured at 90 °C for 30 s before being loaded from a Unimetrics microsyringe into comb-formed slots in a 0.3 mm-thick gel matrix immersed in TBE. The gel solution was prepared as 6.65% propenamide, 0.35% N,N'-methane-*bis*propenamide, 7 M carbonyl diamide (urea), 0.075% fresh ammonium persulphate in 80 cm^3 of TBE and dissolved completely before adding 0.075% by volume N,N,N',N'-tetramethylethane-1,2-diamine (TEMED) to initiate polymerization. The gelling solution was then poured between siliconized glass plates from a 50 cm^3 syringe within 10 min.

Electrophoresis was carried out at 30 W for 2-3 h until the leading dye had migrated about 35 cm from the origin. The plates were then separated leaving the gel adhering to one of them; the gel was overlaid with a sheet of 'Hybond' for its eventual transfer and also to facilitate its removal from the plate, as the 'Hybond' now maintains the gel's integrity. DNA transfer was conducted by means of a 'Novablot' electroblotting apparatus (Pharmacia-LKB) consisting of two graphite slab electrodes between which are stacked a series of 'trans' units. Each trans unit is a convenient section of 'Hybond' and adhering gel sandwiched between three TBE-soaked filter papers, neighbouring trans units being separated by cellulose dialysis membranes. After transfer at 0.5 A, 30 V for 30 min the 'Hybond' was rinsed, dried and baked as before.

2.13. Ligation and subcloning

To subclone smaller fragments from existing clones into plasmid vectors, 10 µg of the cloned DNA was digested with a suitable restriction enzyme which would cleave within the insert but not in the vector sequences; the enzyme was then denatured by heating to 70 °C for 10 min. Where one particular fragment was required for subcloning, it could be isolated by electrophoresis in an agarose gel containing ethidium bromide followed by identification of the desired band under long-wave (366 nm) UV illumination. The method of Dretzen *et al.* (1981) was used to purify the fragment: a small square (9 mm^2 x 0.45 µm) of NA45 DEAE membrane (Schleicher and Schuell) was inserted into a slit created in the gel immediately adjacent to the band, the gel rotated by 90 ° and electrophoresis continued until all DNA in the band had been adsorbed to the membrane which was then removed, rinsed twice in TE and the DNA eluted by incubating at 65 °C for 30 min in a small volume (60 mm^3) TE containing 1 M NaCl. The eluted sample was subsequently spun through a column of Sephadex G50 (Pharmacia-LKB) to remove the NaCl.

In cases where any one of a number of fragments would be acceptable for subcloning, there is no need to isolate specific bands and the digested samples, following denaturation of the restriction enzyme, can be used directly in a ligation reaction.

A 10 µg sample of plasmid vector DNA (pGEM1, Promega Biotech) was cut within its multiple cloning site ('polylinker') by a suitable restriction enzyme to generate cohesive ends complementary to those of the intended insert DNA. After making up to 50 mm^3 with TE and 10 mM Tris pH 8, the linearized plasmid was treated with 0.1 unit calf intestinal phosphatase (Boehringer-Mannheim) for 1 h at 37 °C to remove its terminal phosphate group and thereby minimize the chance of it re-constituting in the ligation. A final concentration of 20 mM 1,2-*bis*[2-aminoethoxy]ethane-N,N,N',N'-tetraethanoic acid (ethyleneglycol-*bis*[β-aminoethylether]-N,N,N'N'-tetraethanoic acid, EGTA; Sigma) was added, the enzymes denatured at 70 °C and the DNA phenol-extracted and ethanol-precipitated before being resuspended in TE at approximately 5 ng mm^{-3}.

Each ligation reaction was set up with approximately equal masses of phosphatased vector and cloned DNA (around 10 ng of each) in a total volume

of 10 mm^3 ligase buffer with 1 unit T4 DNA ligase (Boehringer-Mannheim). This procedure gives a slight excess of insert over vector ends if the former is the shorter of the two, which also helps to minimize self-ligation of the vector. The proportion of self-ligations to intermolecular ligations was monitored *via* control ligations containing linearized plasmid DNA only; it was found that cloning efficiencies approaching 100% could be obtained by this method. The reactions were allowed to proceed at 14 °C overnight; half of each ligated sample, together with appropriate controls of intact and linearized vector DNA was then used to transform competent cells of *E. coli* (see below).

2.14. Transformation

Transformation of *E. coli* with plasmid DNA was accomplished using the strain DIH101, a derivative of HB101 (Boyer and Roulland-Dussoix, 1969). DIH101 was constructed by Dr D. Ish-Horowicz by transposon-mediated integration of part of the F-episome containing the male fertility genes, and is Km^R Ap^S until transformed by a plasmid carrying the β-lactamase gene conferring ampicillin-resistance (Ap^R).

Cells were made competent to take up plasmids by the following adaptation of the Hanahan (1983) method. Single fresh colonies on ψ-agar were used to inoculate 5 cm^3 of ψ-broth and grown at 37 °C to an $O.D._{550}$ of 0.3, then subcultured in 100 cm^3 of ψ-broth to an optimal $O.D._{550}$ of 0.48. The cells were chilled on ice for 5 min before being divided between four centrifuge tubes and spun at 4 000 *g* for 5 min at 4 °C. The pellets were gently resuspended in 9 cm^3 each of TfbI, chilled on ice for 5 min and spun down again at 4 000 *g* as before. The pellets were this time resuspended in 0.9 cm^3 TfbII, chilled on ice for 15 min and divided into 50 mm^3 aliquots before being snap-frozen in liquid N_2 and stored at -70 °C.

To transform cells with plasmid, an aliquot was thawed and immediately left on ice for 10 min. Between 0.1 and 10 ng of plasmid DNA was added and the sample kept on ice for a further 30 min. The cells were then heat-shocked at 42 °C for 90 s and immediately returned to ice for 2 min, during which time plasmids should enter the cells. To allow expression of the plasmid-encoded ampicillin resistance, 800 mm^3 of ψ-broth was added and the transformed cells incubated at 37 °C for 50 min. The cells were then pelleted in a microcentrifuge for 10 s, resuspended in a smaller volume

(100 mm^3) of supernatant and finally plated out by spreading on L-agar supplemented with 100 µg cm^{-3} ampicillin. Following overnight incubation at 37 °C, an efficiency of 10^6 colonies μg^{-1} supercoiled DNA was regarded as satisfactory.

In order to determine whether any of the plasmids contain the desired cloned insert, it is necessary to sample single colonies randomly and make DNA 'minipreps' from them on a trial-and-error basis. The sizes of the plasmids may then be determined by electrophoresis alongside standard size markers, and plate stocks derived from suitable clones used to isolate larger quantities of DNA for use as probes.

2.15. Preparation of plasmid DNA

Small samples of plasmid DNA ('minipreps') were prepared by the boiling method of Holmes and Quigley (1981). This involved growing cells from a single fresh colony in 10 cm^3 of L-broth + ampicillin overnight, from which 1.5 cm^3 was removed and centrifuged in an Eppendorf tube for 20 s. The pellet was resuspended in 200 mm^3 of STET and lysozyme added to 0.9 mg cm^{-3}. After thorough mixing, the suspension was boiled at 100 °C for 40 s and immediately spun in a microcentrifuge for 10 min. The flocculent pellet of bacterial debris was removed with a sterile toothpick, and DNA recovered from the remaining phase by ethanol-precipitation at -70 °C.

Large-scale preparation of plasmid DNA followed the alkaline lysis method described in Maniatis *et al.* (1982). Bacterial pellets from a 400 cm^3 overnight culture in L-broth + ampicillin were resuspended in 9 cm^3 each of 50 mM glucose, 25 mM Tris pH 8, 10 mM EDTA and pooled into one bottle. Lysozyme was added to 4 mg cm^{-3}, and after 10 min at room temperature 40 cm^3 of fresh 0.2 M NaOH, 1% SDS was added with thorough mixing before chilling on ice for 5 min. To this was added 20 cm^3 of solution III, mixing well immediately; after chilling on ice for more than 15 min, 10 cm^3 distilled water was added and the mixture centrifuged at 10 000 *g* for 5 min. The supernatant was filtered to ensure complete removal of cell debris, 0.6 volume propan-2-ol added and the DNA spun down at 10 000 *g* for 5 min. The pellet was dried thoroughly under vacuum and resuspended in 11 cm^3 TE.

After adding 0.6 cm^3 0.2 M EDTA and 0.1 cm^3 of 2 M Tris-base, the

solution was made up to 13.4 cm^3 with TE and 14.80 g CsCl added followed by 1.4 cm^3 of 10 mg cm^{-3} ethidium bromide. The final solution of 1 g cm^{-3} CsCl and 0.9 g cm^{-3} ethidium bromide was centrifuged to equilibrium at 125 000 *g* for more than 40 h and the lower band of closed circular plasmid DNA collected with a siliconized Pasteur pipette. Ethidium bromide was removed by passing the sample through a 4 cm^3 Dowex AG50W-X8 (Bio-Rad) ion exchange resin column equilibrated with 1 M NaCl, 0.01 M EDTA and 0.1 M Tris pH 7.5 prior to use.

2.16. Synthesis of radioactively-labelled probes

Radioactive probes were prepared in one of two ways: either by the method of oligo-labelling (Feinberg and Vogelstein, 1983) or by *in vitro* transcription (Melton *et al.*, 1984). The first of these involves *de novo* synthesis of DNA between oligonucleotide primers which have been annealed to the denatured strand of cloned DNA. The DNA was first linearized and, if possible, the desired fragment purified by elution from DEAE membrane (Dretzen *et al.*, 1981; see 2.13) to increase the intensity of hybridization.

About 100 ng of fragment, together with 30 ng of the appropriate marker DNA, was denatured at 100 °C for 90 s, 3 mm^3 oligo-labelling buffer (OLB) added immediately and the mixture cooled to room temperature before adding 0.1 cm^3 bovine serum albumin, 3 mm^3 DNA polymerase I (Amersham International) and 7.4 MBq α-[^{32}P]dCTP (110 TBq $mmol^{-1}$; Amersham International). The polymerase was found to achieve comparable specific activities to the more expensive Klenow fragment usually recommended, provided that the reaction was not left too long.

After 1.5 h incubation at 14 °C, incorporation was monitored by dissolving excess free nucleotide in 5% trichloroethanoic acid. If satisfactory (>50%), the reaction mixture was spun through a Sephadex G50 column equilibrated in TE and the DNA denatured in 0.25 M NaOH, neutralized with an equal volume of 1 M Tris pH 7.5 and finally added to the hybridization solution in a polythene bag containing one or more 'Hybond' filters (see below).

The second method of labelling generates a radioactive RNA probe by transcription from the cloned DNA, and is only possible for fragments cloned into a plasmid vector containing a suitable transcription promoter. The plasmid

must be linearized in such a way that the promoter retains both *cis* contiguity with the insert and the correct orientation for transcription to proceed into the insert sequence, which should preferably be terminal. For fragments cloned into the *Eco*RI site of pGEM1 this was achieved by digestion with *Sal*I which cleaves downstream in the polylinker.

The labelling reaction comprised 100 ng linear probe DNA, 15 nmol each of ribonucleotides ATP, GTP and UTP, 10 units each RNAse inhibitor and T7 RNA polymerase (Boehringer-Mannheim) and 6 MBq α-[^{32}P]CTP (30 TBq mmol^{-1}, Amersham International) in a total volume of 30 mm^3 transcription buffer. The reaction was allowed to proceed at 37 °C for 45 min before it was monitored for satisfactory incorporation, arrested by heating to 65 °C with the addition of 10 mM EDTA, and added to the hybridization solution together with an appropriate oligo-labelled marker probe.

2.17. Hybridization of filters to radioactively-labelled probes

Membrane filters were wetted with 2 x SSC, placed in a 'Polybadge' polythene bag (Jencons Scientific) and immersed in the hybridization solution which had been pre-warmed to 65 °C. This consisted of 0.3% low-fat dried milk (J. Sainsbury), 0.5 M sodium phosphate buffer pH 7.2, 7% SDS and 1 mM EDTA. Low-fat dried milk is an inexpensive and effective alternative to Denhardt's solution (Johnson *et al.*, 1984). Radioactive probe DNA was added to the bag, mixed evenly through the solution and allowed to incubate overnight at 65 °C.

The radioactive solution was then discarded and the filters washed in 1 mM EDTA, 40 mM sodium phosphate buffer pH 7.2, 1% SDS twice at 57 °C for 30 min, then rinsed in a third batch of the same solution at 57 °C before being sealed in thin polythene bags and overlaid with a sheet of X-ray film in an autoradiographic cassette. After an appropriate time the films were developed for analysis of the bands corresponding to the DNA fragments.

2.18. Removal of probe DNA from membranes

The binding of DNA fragments to robust nylon membranes such as 'Hybond' after baking is sufficiently tenacious to provide a long-term record of their electrophoretic mobility, to which successive probes may be hybridized

and removed. Treatment with 0.4 M NaOH at 42 °C for 30 min followed by neutralization in 1 M Tris pH 7.5, 1.5 M NaCl is most efficient at stripping away probe DNA, but has also been known to detach all fragments under 0.4 kb from the membrane.[2] In cases where small fragments were of interest, therefore, a milder stripping procedure was substituted using 96% methanamide (formamide), 10 mM Tris pH 8, 10 mM EDTA at 80 °C for 30 min and rinsing afterwards in SSPE.[3]

2.19. DNA used as probes

A series of overlapping phage clones from the *achaete-scute* complex was provided by Dr J. Modolell (Campuzano *et al.,* 1985) and sub-clones into pUC8 (Vieira and Messing, 1982) constructed by Robin Beech (Beech, 1987; Beech and Leigh Brown, 1989) for use as probes. Additional fragments were purified from these or from the original phage clones as required; the locations of all probes on the molecular map of *ac-sc* are given in Fig. 6.

Genomic fragments of DNA cloned from the *ry-Ace* region into λCharon4 vectors from the Canton-S library of Maniatis *et al.* (1978) were kindly provided by Dr P. Spierer (Spierer *et al.,* 1983; Hall *et al.,* 1983). The phage DNA was packaged *in vitro* and DNA maxipreps derived from four clones corresponding to the genes mapped in this region: 2842, 2837, 2827 and 2118. From these DNA samples, fragments of a suitable size (2-3 kb) were subcloned into pGEM1 which contains the phage T7 promoter, a useful ally in the radioactive labelling process (see 2.16).

In addition, a 4.6 kb *Eco*RI fragment including the entire *rosy* gene sequence and cloned into a pBR322 vector (Rushlow *et al.,* 1984; Keith *et al.,* 1987) was amplified and plasmid DNA prepared from the cells used to subclone the 3.8 kb *Bgl*II fragment from within the insert into pGEM1 *via* the *Bam*HI site in the polylinker; such a process is possible because the enzymes *Bam*HI and *Bgl*II cleave to produce cohesive ends of like sequence. This

---

[2] J.A. Lyon, 1985. Genetics 4 Project, University of Edinburgh.

[3] D.E. Graham, 1984. NEN Product News 4,1.

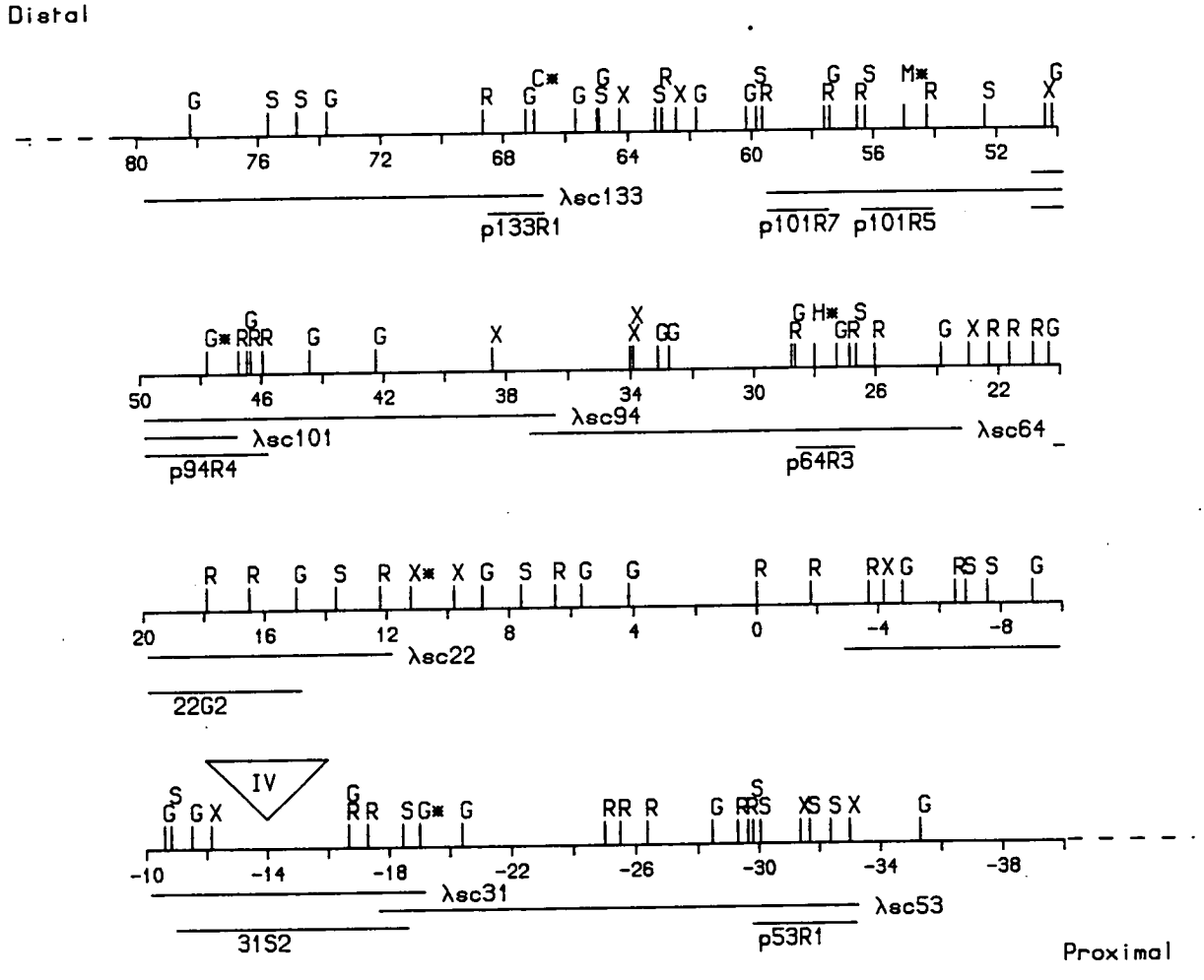
Figure 6

Summary restriction map of the *achaete-scute* complex (ordinates of Campuzano *et al.,* 1985). Below the map are shown the positions of some of the λ phage clones from the region, and below these the various fragments excised from them for use as probes (see 3.1 in text). Those prefixed by 'p' were also subcloned into plasmids. Polymorphic restriction sites are denoted with an asterisk (*); note that while all sites are shown for the six-cutter enzymes, only polymorphic four-cutter sites have been included. C=*Cfo*I; M=*Msp*I; H=*Hae*III. G=*Bgl*II; S=*Sal*I; R=*Eco*RI; X=*Xba*I. The approximate size and position of insertion IV is also given.

Opposite page 33

Fig.6

Molecular map of the *ac-sc* complex



clone, named pRA-G3, was the principal one used to probe the *rosy* gene region although another, 1.7 kb *Bam*HI/*Eco*RI fragment was later subcloned (pRA-BR10) and used to investigate the 3' end of the gene. A 1.9 kb *Sal*I fragment (42-S3) was purified from the phage 2842 DNA specifically to investigate a known polymorphism in the vicinity of the /S12 gene (Aquadro *et al.*, 1988). The molecular map of the *ry* region, showing the positions of all these clones and fragments used as probes is summarized in Fig. 7.

2.20. Analysis of genomic DNA

The screening of the two genomic regions used methods very similar to those of Kreitman and Aguadé (1986). Each line of genomic DNA was analysed by digestion with a range of restriction enzymes, the digested samples divided and fragments separated by the complementary techniques of agarose- and *poly*propenamide gel electrophoresis, transferred and fastened to nylon membranes, hybridized to a homologous radioactively-labelled probe and eventually visualized by autoradiography. The principal enzymes used were *Alu*I, *Cfo*I, *Hae*III, *Mbo*I, *Msp*I, *Rsa*I and *Taq*I all of which recognize 4-base-pair sequences; *Bam*HI, *Bgl*II, *Sal*I and *Xba*I which recognize 6-base-pair sequences, and *Ban*II which recognizes a sequence of six base-pairs, two of which are flexible (A or G, and T or C recognized). The latter therefore cleaves more frequently than conventional six-cutters, but will fail to detect any transitions at these two flexible positions in the site.

Following successful analysis of one genomic sequence, the membranes were stripped of the probe (see 2.18) and subsequently re-hybridized with another labelled fragment from elsewhere in the region. In this manner, large tracts of DNA from many lines of flies could be screened relatively rapidly. Initially, small samples of 10-16 lines were digested in turn with the entire range of enzymes as a preliminary screen for variation. Any polymorphisms detected in samples of this size are likely to be at high frequency; such variants were then investigated further by digesting larger samples of up to 72 lines with the relevant restriction enzyme. This strategy is economical on both enzyme and genomic DNA.

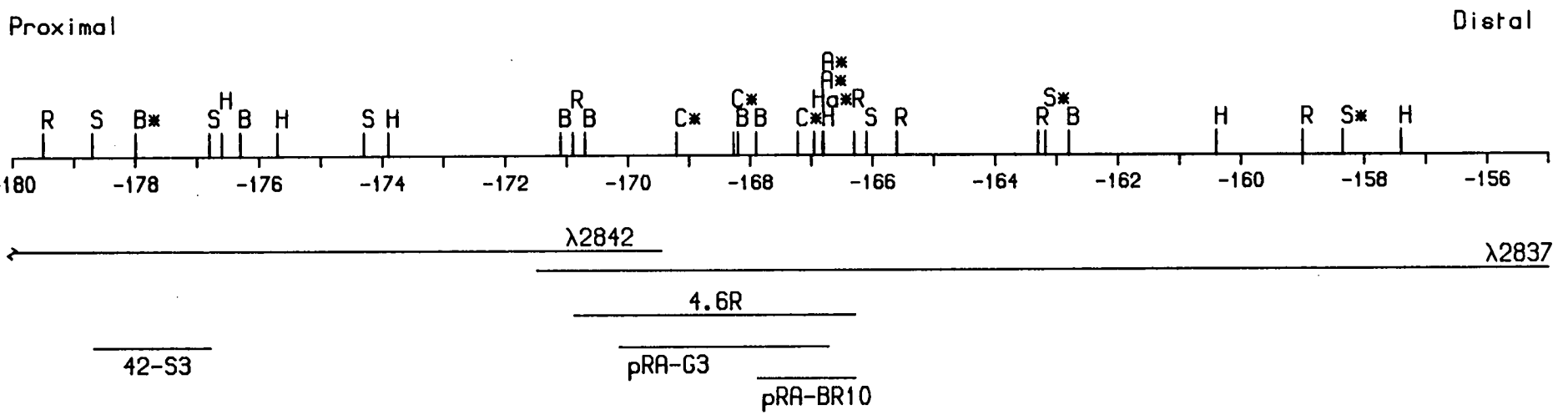
Figure 7

Summary restriction map of the 30 kb of chromosome III encompassing the *rosy* and *rS12* genes (ordinates of Bender *et al.,* 1983), showing the positions of λ phage clones and fragments excised or subcloned from them. Fragment 4.6R is an *Eco*RI fragment subcloned into pBR322 (Rushlow *et al.,* 1984) from which pRA-G3 and pRA-BR10 were subcloned into pGEM1. Fragment 42-S3 was excised directly from λ2842. Polymorphic restriction sites are denoted with asterisks (*); only the polymorphic sites are shown for the four-cutter enzymes. A=*Alu*I; C=*Cfo*I; Ha=*Hae*III. B=*Bam*HI; H=*Hin*dIII; R=*Eco*RI; S=*Sal*I.

Opposite page 35

Fig.7

Molecular map of the /S12-*ry* region



CHAPTER 3

RESULTS

The correspondence between the numbering of fly lines used here and that of Beech (1987) and published work (Beech and Leigh Brown, 1989; Macpherson *et al.*, in preparation) is given in Appendix III; 17 additional X chromosome lines were harvested of male flies and DNA samples prepared from them. Of the 72 third-chromosome lines, 36 did not yield sufficient numbers of homozygotes for DNA preparation; these lines were made hemizygous for a deletion as described in Methods (2.7) and are identified with asterisks in Appendix III. That the deficiency stock had been successfully converted to P cytotype beforehand was confirmed by reciprocal testing: on crossing to a strong P strain of sn^w, only one sn^e revertant was recovered out of 236 F_2 progeny; the converse cross to a known M strain of sn^w produced high numbers (20%) of sn^+ and sn^e F_2 revertants. A similar result was obtained when the *TM6B/TM3* stock used to introduce the *TM3* balancer was tested (34% reversion in F_2 on crossing to M).

Both the extracted X and third-chromosomes are prevented from crossing-over with their homologues throughout their time in the laboratory, and are therefore assumed to represent the wild state of these chromosomes. *TM6B* has been described as 'the best third-chromosome balancer currently available'[1] while any breakdown of the attached-X system which could lead to recombination of the wild X chromosome would be accompanied by the loss of the *yellow* phenotype of the females.

3.1. The *ac-sc* region

The screening of the *achaete-scute* complex for variation had been initiated previously and revealed a series of insertion-deletion and restriction-site polymorphisms (Beech, 1987). Analysis of all those whose frequency exceeded 0.05, *i.e.* with the rarer allele represented more than twice in the sample of 44 lines, showed unusually high linkage disequilibrium to exist between most pairs of variants over the entire range of distances (Beech

[1] L. Craymer, 1984. Drosophila Information Service 60, 234.

and Leigh Brown, 1989). Four of these polymorphisms (G-19, G48, X11 and insertion IV) were further elucidated during the course of this work (Figs 8-13.) The putative insertion-deletion event designated 'A' was, however, not scorable reliably from the data presented here (Fig. 14).

Polymorphic restriction sites are shown with an asterisk on the summary restriction map of the *ac-sc* complex (Fig. 6), together with insertion IV. Although shown singly, it is possible that IV is a multiple event. The ordinates on the map are those given by Campuzano *et al.* (1985) from which the approximate positions of six-cutter polymorphisms could be deduced. The use of four-cutter enzymes revealed a further three site polymorphisms, assigned ordinates as the nearest integer to the centre of the probe with which they were detected; these positions are therefore subject to possible errors of around 2 kb.

Nomenclature for the subcloned probes followed the convention adopted by Beech (1987): 'pASC' followed by the number of the recombinant phage, the enzyme used in subcloning and the number of the fragment in the phage in order from proximal to distal. Thus the plasmid pASC133R1 contains the first *Eco*RI fragment from the proximal end of λsc133, and pASC101R7 the seventh *Eco*RI fragment from the proximal end of λsc101. In Fig. 6 'pASC' is abbreviated to 'p' for simplicity, and omitted altogether for fragments isolated without subcloning.

3.1.1. M55 polymorphism

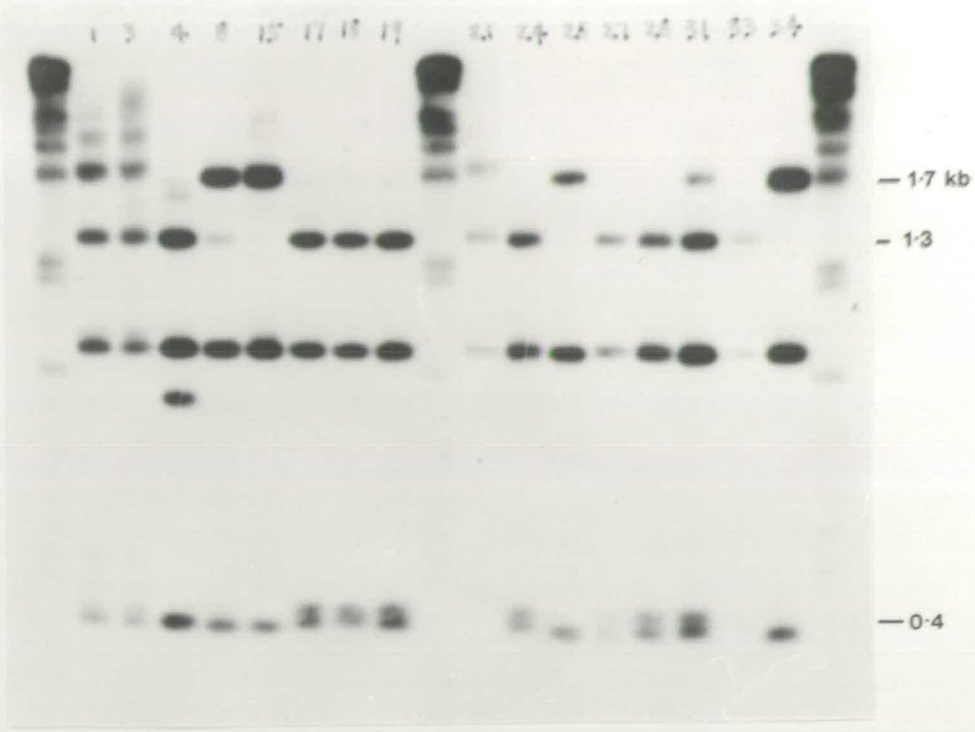
The probe pASC101R5 revealed variation in lines digested with *Msp*I (Fig. 8) whereby fragments of 1.3 and 0.4 kb combine into a single one of 1.7 kb at a frequency of 0.16. This observation is consistent with the abolition of a *Msp*I recognition site at around +55 kb on the molecular map. An additional variant band of around 0.7 kb was detected in line 4 using this combination of enzyme and probe, but this appears to be unique in the sample and has been interpreted as an insertion event, since pASC101R5 showed line 4 to be polymorphic also for enzymes *Mbo*I and *Rsa*I and the number of bands suggests an increase in the total amount of DNA hybridizing to the probe.

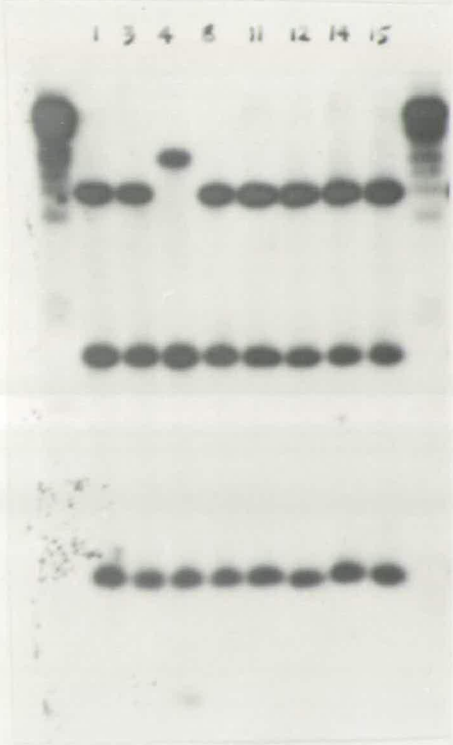
Figure 8

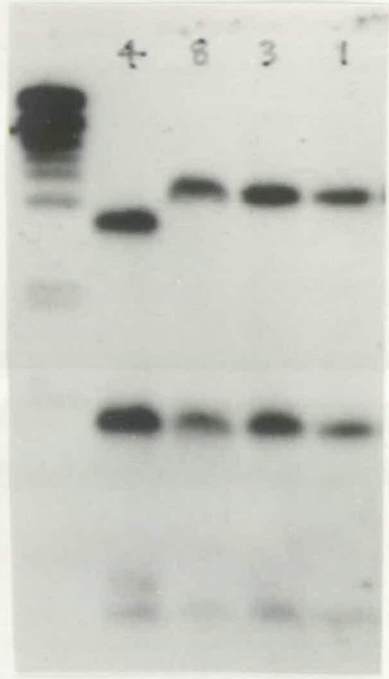
(Top) DNA digested with *Msp*I and probed with pASC101R5 (see 3.1.1), showing the M55 polymorphism and also the extra band in line 4. The latter can be attributed to a sequence-length change since variant bands are also observed in this line when the DNA is digested with other enzymes such as *Mbo*I (bottom left) and *Rsa*I (bottom right). Marker DNA is λ*Pst*I in all cases.

Opposite page 38

Fig. 8







3.1.2. C67 polymorphism

The probe pASC133R1 hybridized only to a single fragment in lines digested with *Cfo*I, indicating that no site for this enzyme exists within the probed sequence of 1.8 kb (Fig. 9). However, the size of this fragment varied between 2.1 and over 4 kb at a frequency of 0.26, consistent with the loss of a *Cfo*I site immediately flanking the probed region. This was given the ordinate +67 kb, the lower estimate of distance from the other polymorphic sites.

3.1.3. H28 polymorphism

Lines digested with *Hae*III and probed with pASC64R3 showed variation in which two fragments of around 750 and 230 bp combined into a single fragment in six out of 42 lines. This difference only showed up reliably on *poly*propenamide gels (Fig. 10), and was attributed to a polymorphic *Hae*III site at approximately +28 kb on the map. With this probe, line 37 differed from the rest when several enzymes were used, indicating a probable DNA length variant in the probed region; again, however, this difference was unique in the sample.

3.1.4. Six-cutter polymorphisms and insertion IV

The polymorphic *Bgl*II site at +47.8 kb (G48) was scored by running samples digested with this enzyme on 0.7% agarose gels and probing the fragments with the insert from pASC94R4 (Fig. 11); the frequency of occurrence of this site was 0.36. Another variant, unique in the sample, shows up as a band of around 10 kb. To investigate the other *Bgl*II polymorphism at −19.1 kb (G−19) and insertion IV together, a 7.6 kb *Sal*I fragment from λsc31 (31S2) was purified as described in Methods (2.13), oligolabelled and hybridized to *Bgl*II-digested samples electrophoresed on 0.8% agarose gels prior to transfer (Fig. 12). Only one further line, 138, scored positive for insertion IV while the final frequency of the G−19 site was 0.43.

The *Xba*I variant at +11.2 kb reduces the size of a 13.1 kb fragment only slightly, to 11.7 kb; samples digested with this enzyme were therefore separated in very low concentration (0.3%) agarose gels. After transfer the fragments were hybridized with a 5.5 kb *Bgl*II fragment isolated from λsc22 (22G2); the final frequency of the site was found to be 0.34 (Fig. 13).

Figure 9

DNA digested with *Cfo*I and probed with pASC133R1 (see 3.1.2); the higher band represents polymorphism C67. Marker DNA is λ *Pst*I.

Opposite page 40

Fig. 9

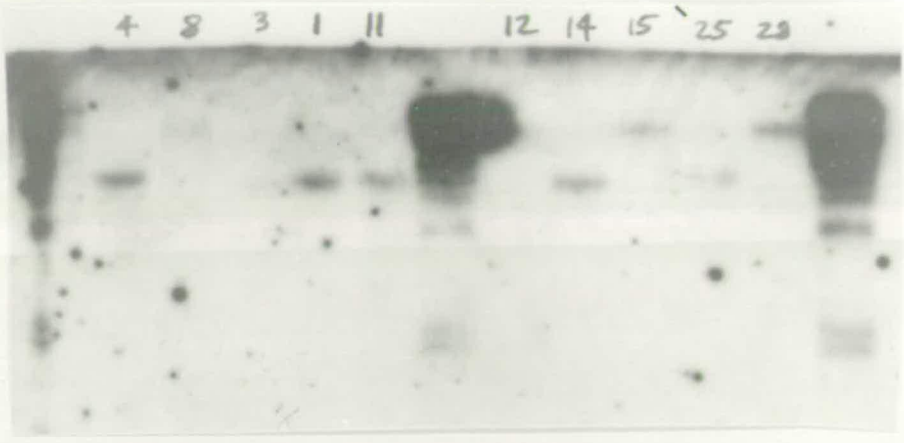
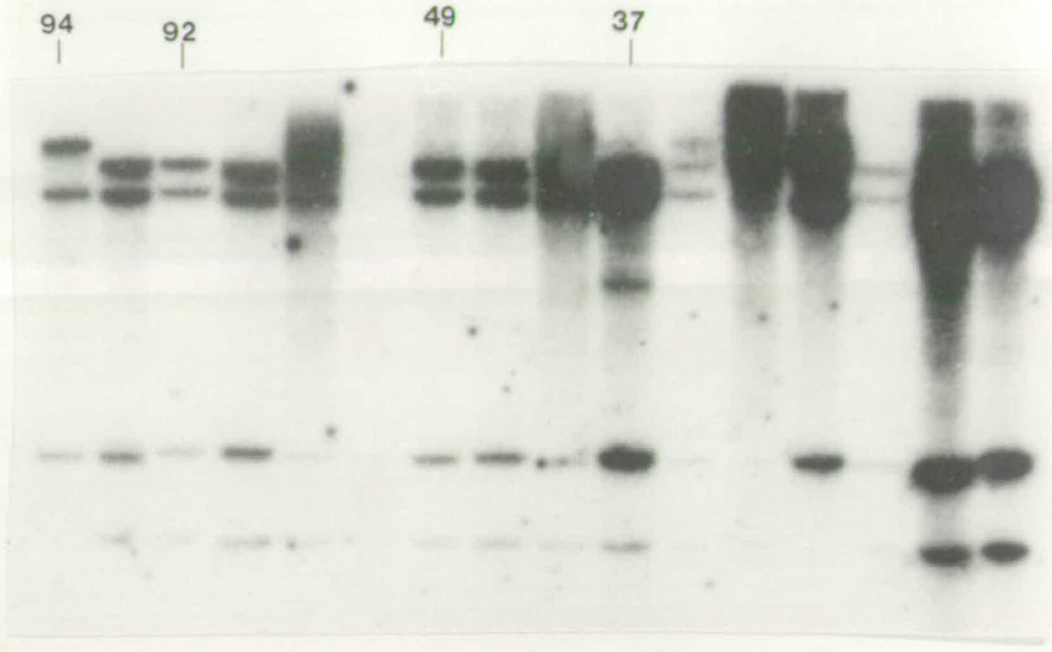


Figure 10

(Top) DNA digested with *Hae*III, electrophoresed in a *poly* propenamide gel and probed with pASC64R3 (see 3.1.3). Line 94 displays the higher band representing the H28 polymorphism. Line 37 shows extra bands both with this enzyme and when *Taq*I is used (bottom), suggesting that an insertion is present in this line.

Opposite page 41

Fig. 10



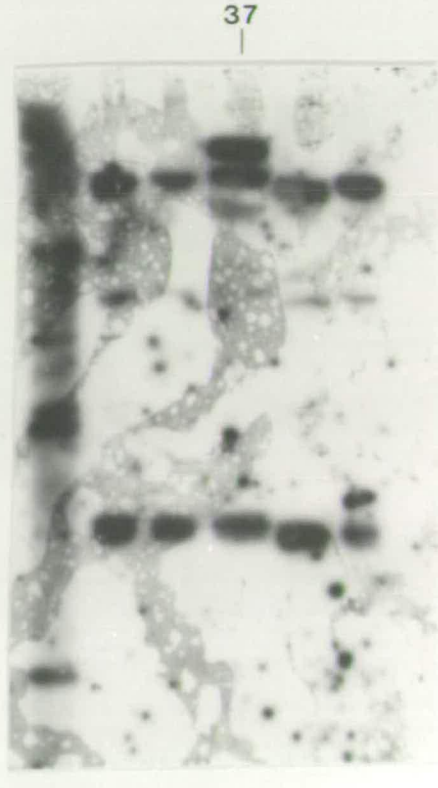


Figure 11

DNA digested with *Bgl*II and probed with pASC94R4, showing the G48 polymorphism (see 3.1.4). The higher band in the fifth track from the right of the gel is a different variant unique to that line. The intense smear in the central track is λ*Pst*I marker DNA.

Opposite page 42

Fig. 11

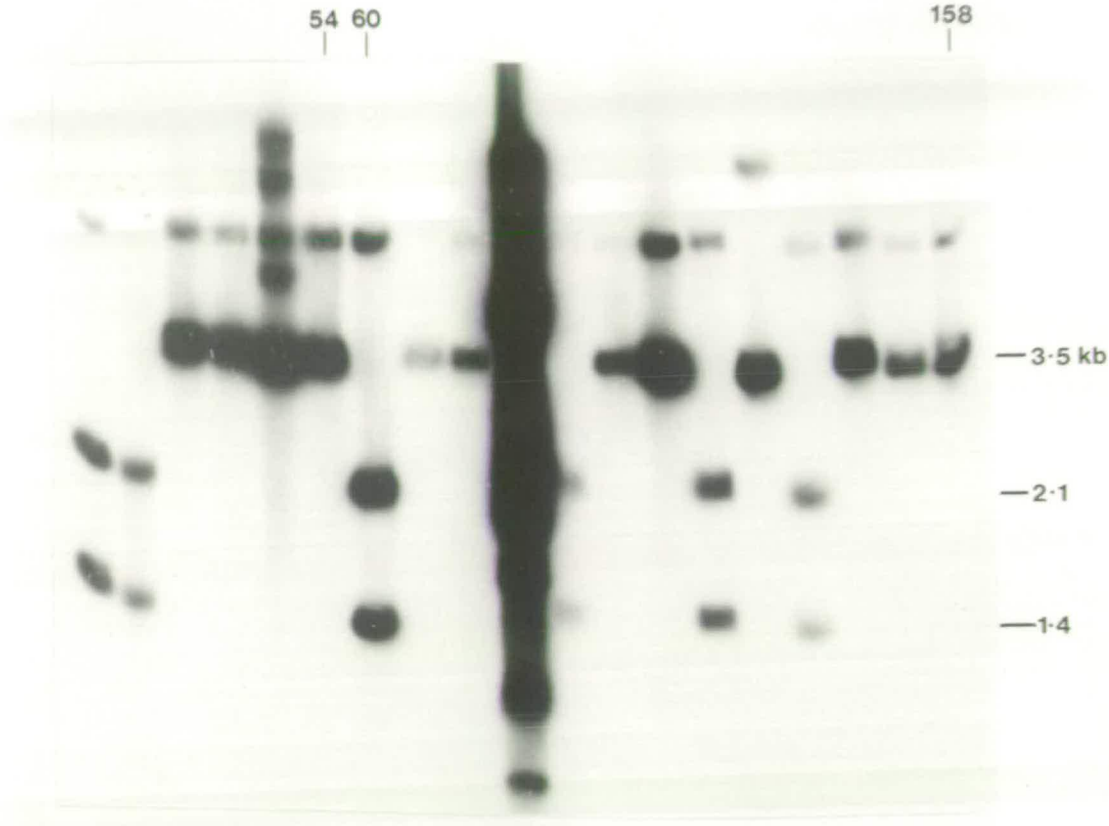


Figure 12

DNA digested with *Bgl*II and probed with sc31S2 (see Fig. 6 and 3.1.4 in text). Insertion IV occurs within the 6.8 kb fragment and hence can be scored independently of the G-19 restriction-site polymorphism in the adjacent 5 kb fragment.

Opposite page 43

Fig.12

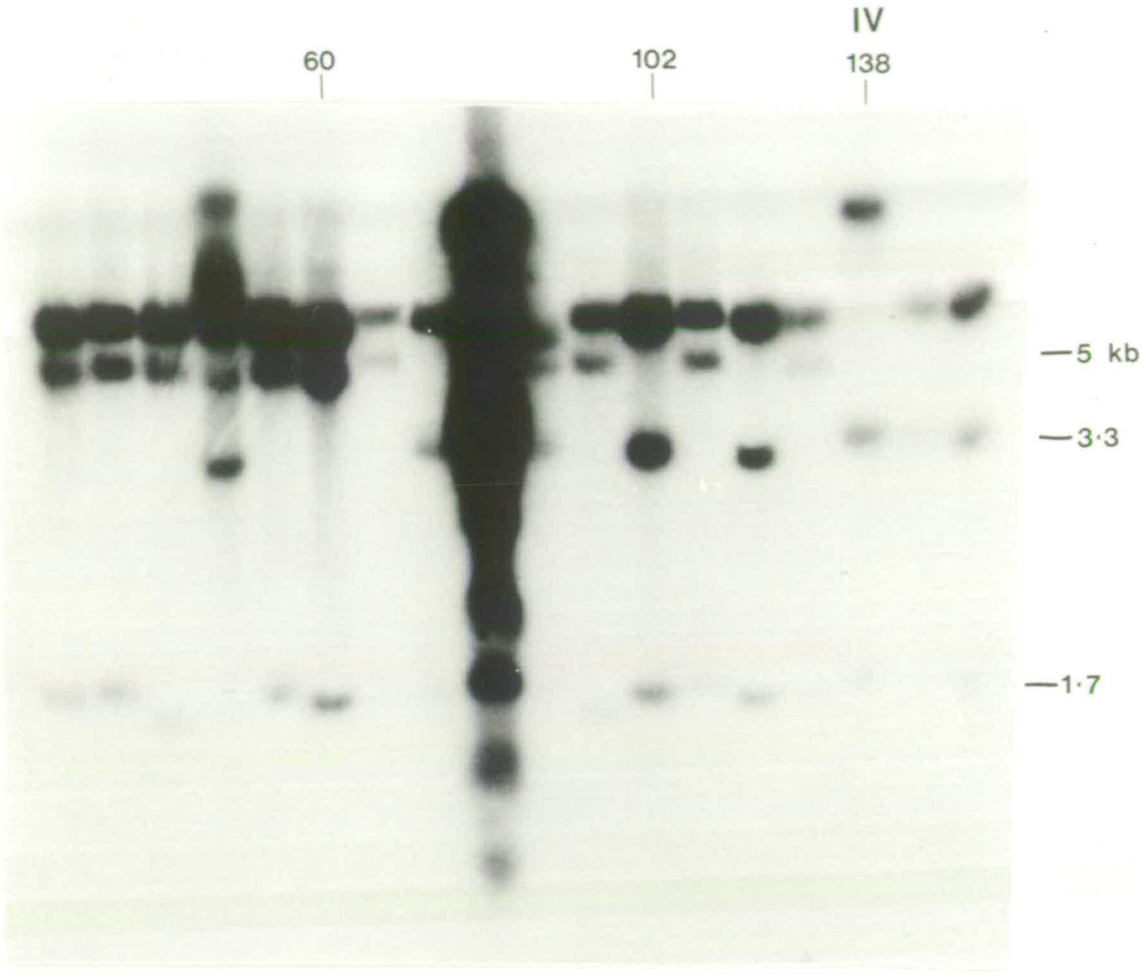


Figure 13

The X11 polymorphism (see 3.1.4) in DNA samples digested with *Xba*I, electrophoresed in a 0.3% agarose gel and probed with sc22G2 (see Fig. 6).

Opposite page 44

Fig. 13

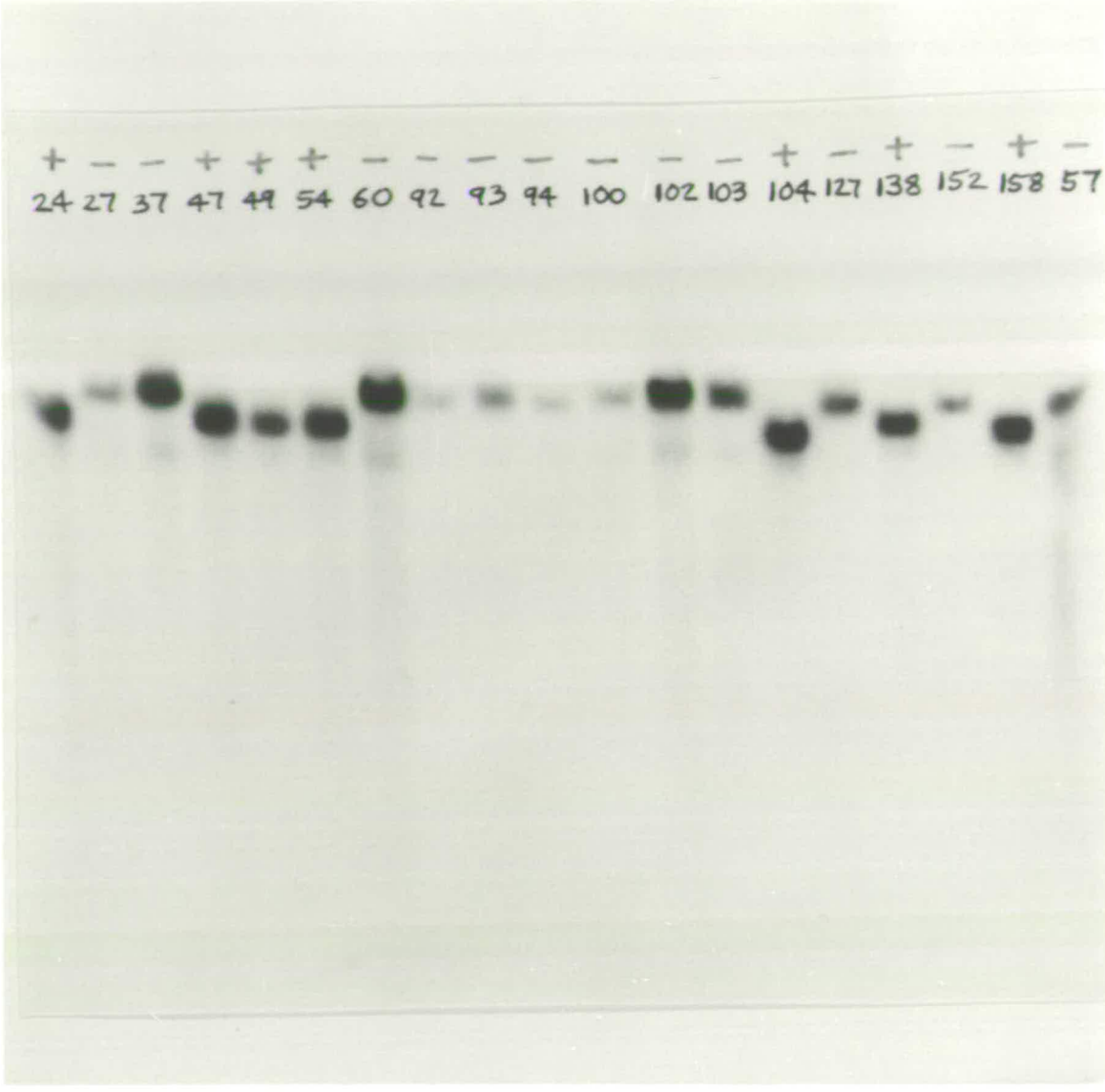
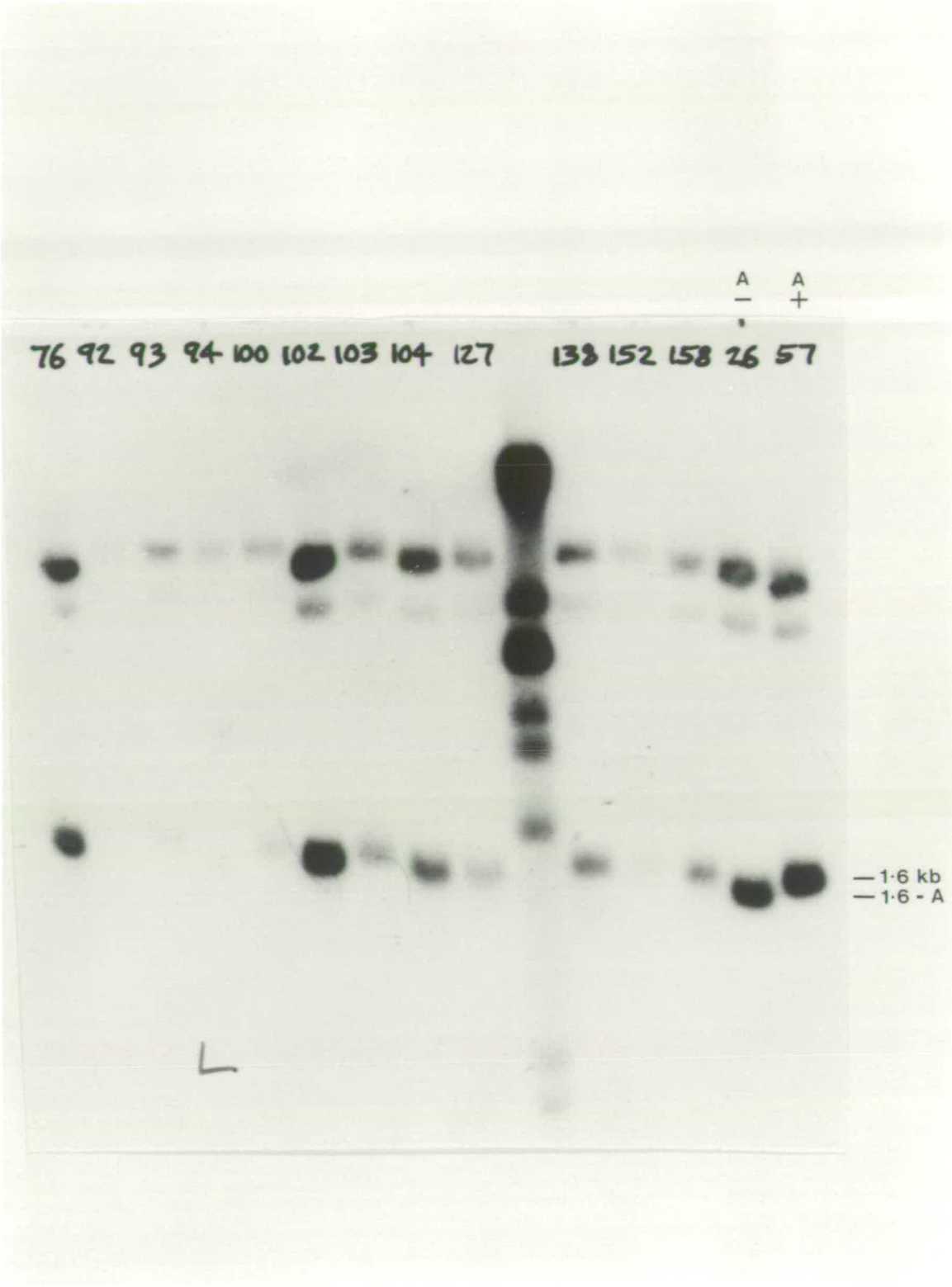


Figure 14

DNA digested with *Bgl*II, electrophoresed in a 1.2% agarose gel alongside size marker DNA (λ*Pst*I) and probed with a 3.3 kb *Sal*I fragment (63.1–59.8 kb on the map) excised from λsc112 (not shown in Fig. 6). The variation in mobility of the 1.6 kb fragment is probably due to the putative small insertion 'A' described in Beech and Leigh Brown (1989) but the fragments did not resolve sufficiently for this to be scored with confidence.

Opposite page 45

Fig. 14



The fragment sizes and frequencies for the seven polymorphisms in the *ac-sc* region are listed in Table A1 of Appendix IV.

3.2. The *ry-Ace* region

A summary map of the entire *ry-Ace* region is given in Fig. 3 showing the positions of genes and transcripts (adapted from Hall *et al.,* 1983; Bender *et al.,* 1983). However, all polymorphic events described here were within 20 kb of the *ry* gene, and are therefore shown on a magnified map of this portion (Fig. 7). The 4.6 kb *Eco*RI fragment used to construct a probe to the *ry* gene is the same one sequenced by Keith *et al.* (1987) so that the polymorphisms detected could in most cases be localized with reasonable certainty to specific base changes. Sequence data from the region could also be consulted *via* the computer database GENBANK using the 'Mapsort' program developed by the University of Wisconsin Genetics Computer Group (Devereux *et al.,* 1984). Nomenclature of the subcloned probes used the prefix pRA, followed by the symbol(s) of the enzyme sites delimiting the insert, then the chronological number of the recombinant colony from which it was derived. The clone pRA-G3, extending between the *Bgl*II sites at -166.8 and -170.2 kb, in fact proved to be the most useful probe for the screening of this region, although a technical problem prevented the insert from being re-isolated in its original form: as it is a *Bgl*II fragment cloned into a *Bam*HI site, neither of these enzymes now recognize its junctions. For purposes of oligo-labelling, therefore, the plasmid was digested with *Eco*RI and *Hin*dIII having first established that the insert was in the correct orientation such that 90% of it would be excised by this procedure. Alternatively, the T7 promoter in the vector was used to generate an RNA transcript for use as a probe (see 2.16).

3.2.1. H-167 polymorphism

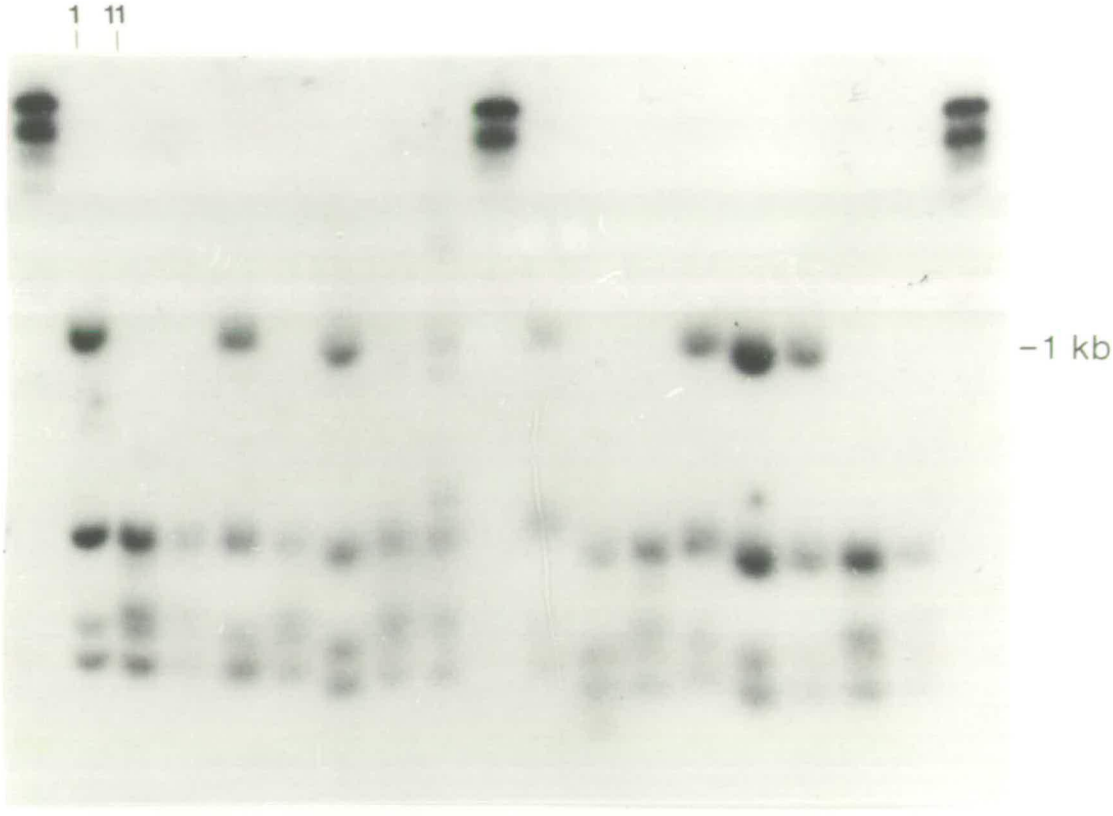
Third-chromosome lines digested with *Hae*III display variation in a 1 kb fragment when probed with pRA-G3, as resolved by both agarose and *poly*propenamide gels (Fig. 15). When the sizes of the fragments are considered in relation to the known sequence, it appears that the variant is an A→C transversion at -167.0 kb, creating a site for *Hae*III recognition cleaving the 1 001-base-pair fragment into daughters of 528 and 473 bp. This position is the only one within the parent fragment where a single nucleotide change to a *Hae*III site would generate fragments of the sizes observed; although the

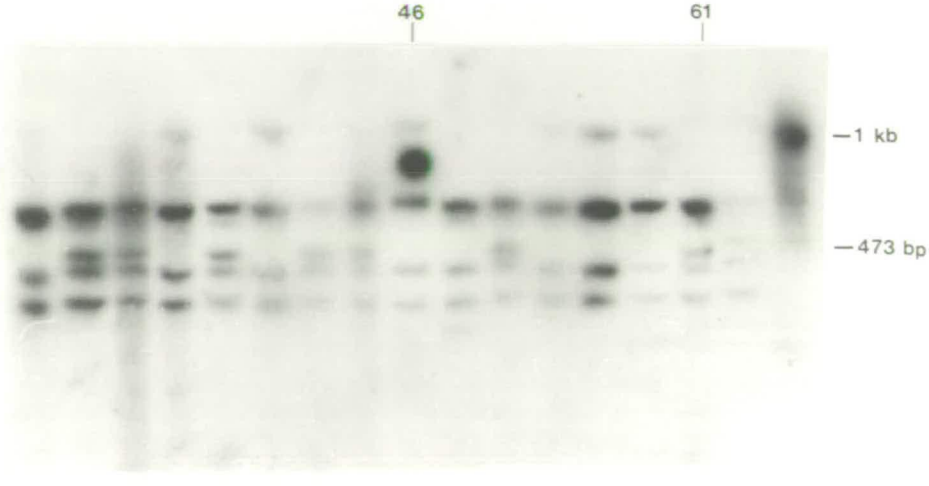
Figure 15

DNA digested with *Hae*III, electrophoresed in 1.8% agarose and probed with pRA-G3 insert DNA (top) or electrophoresed in 7% *poly* propenamide and probed with pRA-G3-derived RNA transcript (bottom). The H-167 restriction-site polymorphism (see 3.2.1) shows up clearly in both cases.

Opposite page 47

Fig. 15





adenine residue is specified in the published sequence, the cytosine allele allowing recognition and cleavage of the site is slightly the more common in my sample at a frequency of 0.55.

3.2.2. A–167 polymorphisms

The same probe revealed variation in lines digested with *Alu*I, involving a small fragment appearing as a low intensity band on the autoradiographs of *poly*propenamide gel blots (Fig. 16). This difference is likely to be due to the abolition of an *Alu*I site increasing the size of the 250 bp fragment by 15 bp; the frequency of this variant was 0.44 in the lines scored, but several lines proved impossible to score owing to the poor intensity of hybridization of such small fragments. In five lines a different variant band is observed, this time replacing the 250 bp fragment with one of around 350 bp. This can be most easily explained by the occurrence of two mutational events in these lines, namely the abolition of both the abovementioned *Alu*I site and the next most distal one 15 bp away. As the next *Alu*I site cleaved would then be a further 94 bp away on the distal side according to the published sequence, such a double event would increase the length of the 250 bp fragment by a total of 109 bp.

3.2.3. *Cfo*I polymorphisms: fine-scale variation

The same pRA-G3 probe detected an apparent polymorphism of frequency 0.16 in lines digested with *Cfo*I (Fig. 17). The extra band immediately above the 326 bp monomorphic band appears slightly higher in line 34 than in the other variant lines, but still lies within the conceivable range of mobilities for a given fragment size. Nevertheless, the merging of bands higher up on the gel makes it unclear which of the larger fragments has been cleaved to produce the novel fragment, so a selection of *Cfo*I-digested lines was electrophoresed in lower concentration (5%) *poly*propenamide gels for a longer period of time in an attempt to distribute the fragments more evenly over the gel.

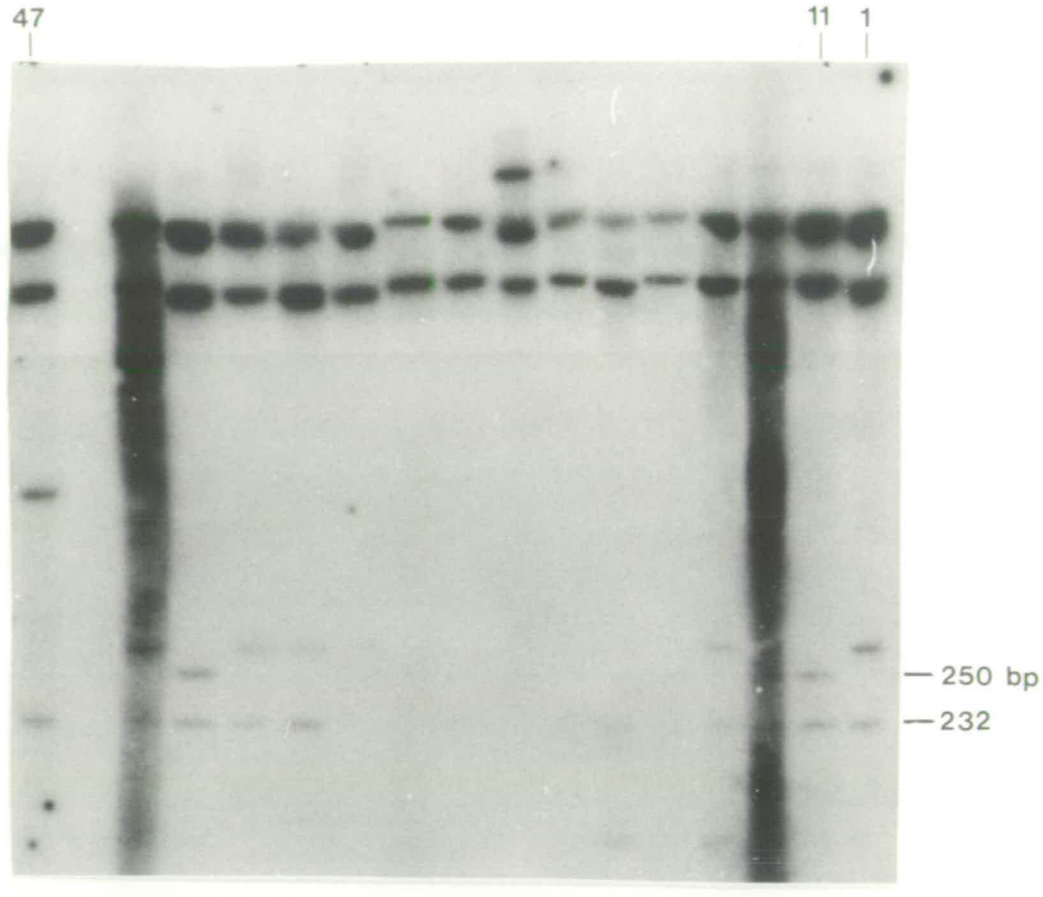
The result, shown in Fig. 18, reveals a startling amount of previously hidden fine-scale variation. The original 'polymorphism' resolves into at least three different ones, of which two can be ascribed to specific single-base substitutions leading to the cleavage of the 507 and 469 bp fragments

Figure 16

DNA digested with *Alu*I, probed with pRA-G3 (top) and pRA-BR10 (bottom). These probes give essentially the same pattern of bands but the latter hybridizes more intensely to the fragments of interest as it is homologous to them over a greater proportion of its length. The first polymorphism detected (A-167) involves only a 15 bp change in fragment length but in lines 47 and 74 the variant band is observed much higher up on the gel, most probably due to the absence of two adjacent *Alu*I sites in these lines (see 3.2.2).

Opposite page 49

Fig. 16



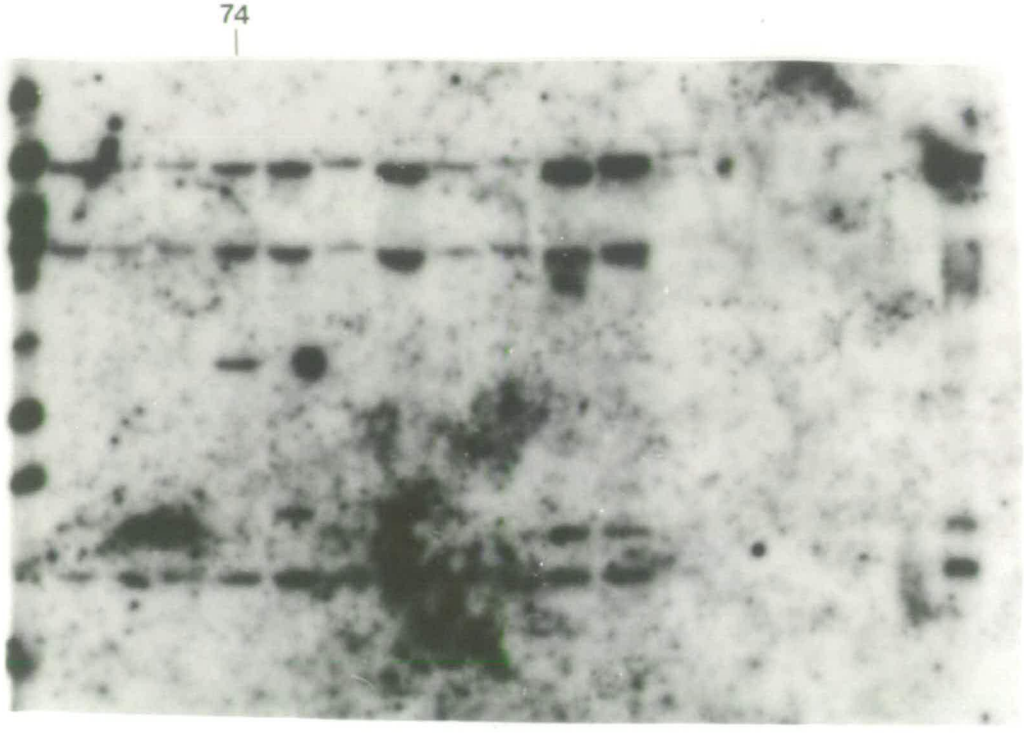


Figure 17

*Cfo*I-digested DNA separated in a gel of 7% *poly*propenamide and probed with pRA-G3, showing the serious pooling effect which can be a feature of electrophoretic assays. The three variant bands appearing in the lines shown were initially thought to represent the same polymorphic event until resolved by more protracted electrophoresis (see 3.2.3 and Fig. 18).

Opposite page 50

Fig. 17

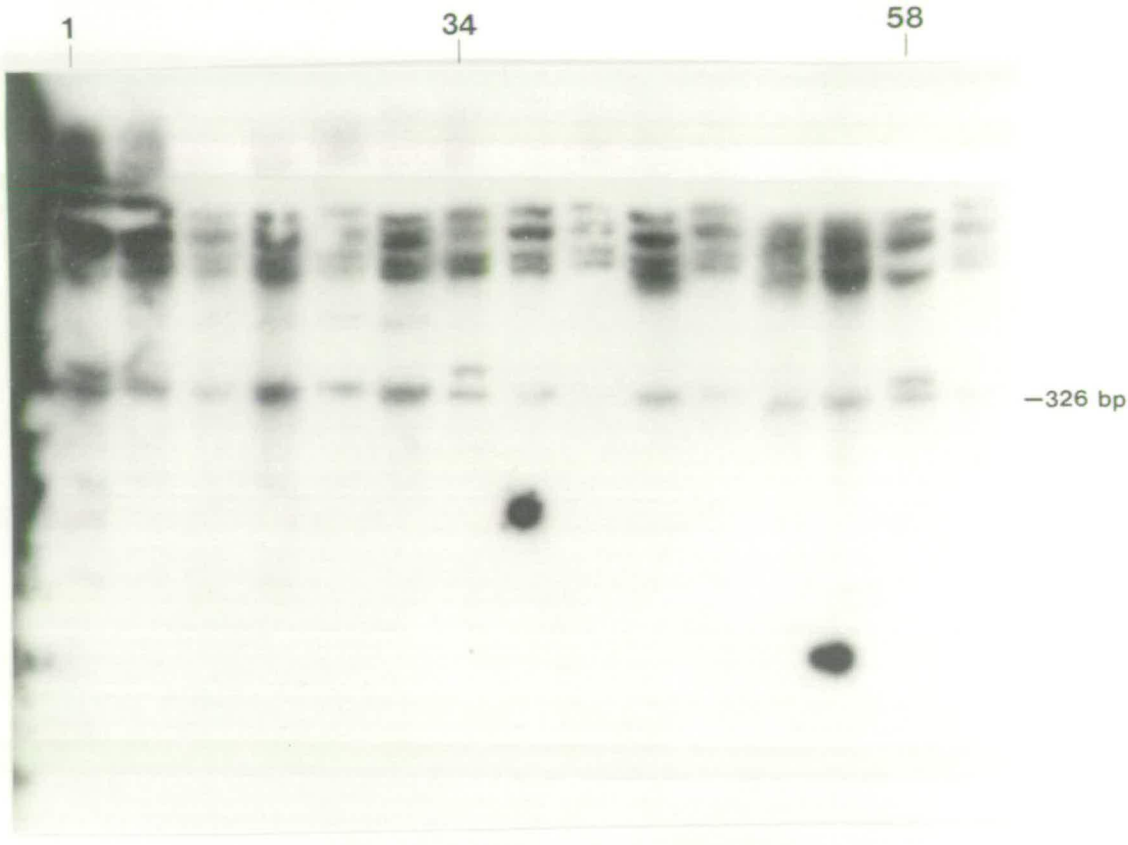
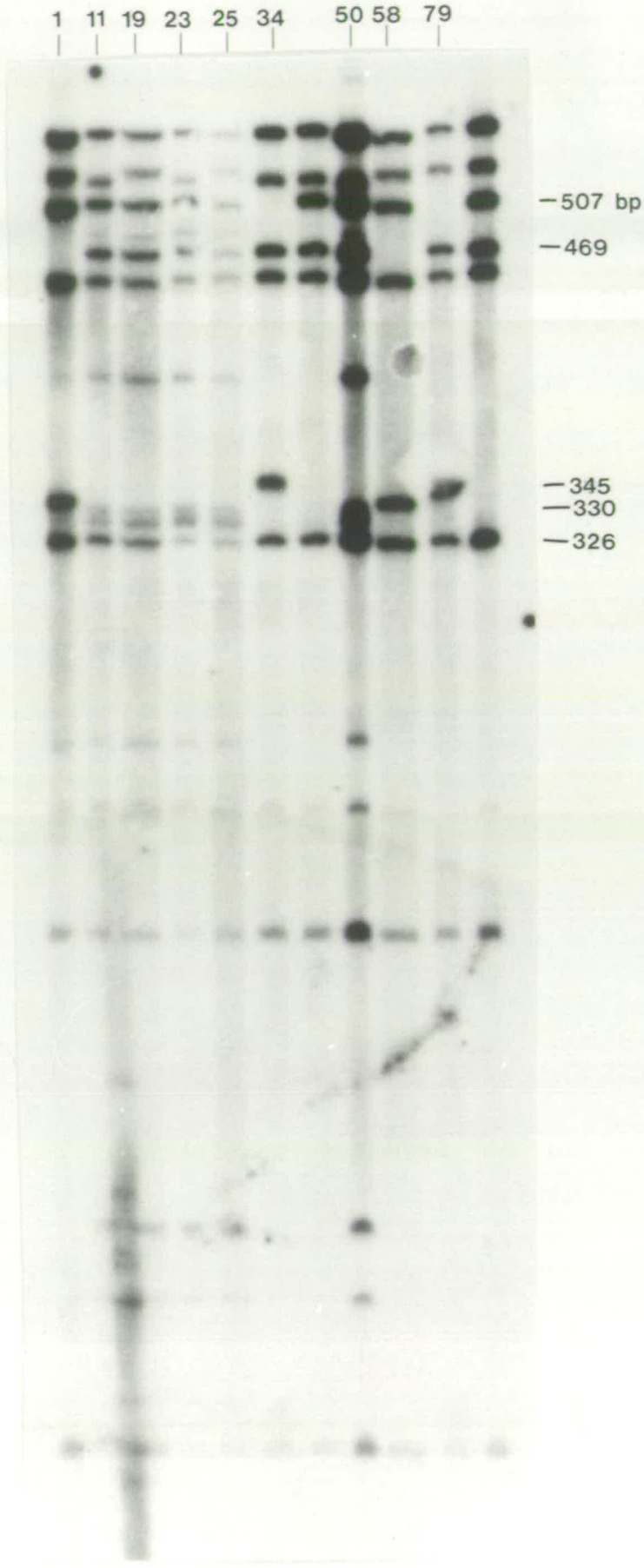


Figure 18

*Cfo*I-digested DNA separated in a low concentration (5%) *poly* propenamide gel over a long period of time (>4 h) to identify fragments which could not be resolved fully in a 7% gel (see Fig. 17). Three distinct variants, all of which appeared similar under conventional conditions, can be discerned here: the restriction-site polymorphisms C-168 (lines 34, 79) and C-169 (lines 1, 58) plus a curious pattern of bands in line 50 (also evident to a lesser degree in 11, 19, 23 and 25) not consistent with any single site- or sequence-length mutation but which also does not seem characteristic of partial digestion. In addition, there is a striking variation in mobility among many of the upper bands over and above that expected from loading artefacts, such as might be produced by small insertion/deletion events (see 3.2.3 in text).

Opposite page 51

Fig. 18



respectively. The variation characteristic of line 50 is less easy to interpret; even an insertion event would not account for all the extra bands present. Lines 1, 11, 12 and 19 have been assigned to the same variant class as 50, but this was not used in the analysis as further work will be required to confirm its exact nature.

The most striking characteristic of the photograph in Fig. 18 is the extensive minute variation in the positions of the upper bands, which at this level of resolution cannot be easily discounted as electrophoretic 'noise' since the distances between bands within tracks is also variable. If confirmed as real, these minor changes would have to be explained as oligonucleotide insertions or deletions which, since some of the fragments involved are known to lie within coding sequences (Keith *et al.*, 1987) would certainly merit further investigation.

All subsequent electrophoresis of *Cfo*I-digested lines was carried out in gels of 5% *poly*propenamide to secure satisfactory resolution of this fine-scale variation. In one of these, a third variant representing a restriction-site polymorphism was uncovered: the 530 bp fragment is lost and another novel fragment appears (Fig. 19.) Minor variation in fragment size is again evident in the upper bands, but in this case the bands are distorted and it could be simply due to uneven voltage gradients across the gel. On the basis of sequence information the three polymorphic restriction sites were given the ordinates -169.3, -168.3 and -167.3 kb in Fig. 7.

3.2.4. Six-cutter polymorphisms

The discovery by Aquadro *et al.* (1988) of three restriction-site variants flanking both sides of the *rosy* gene in a sample from the same NC population prompted a search for these polymorphisms in this sample. The *Bam*HI site on the distal side of *ry* and *l*S12 was scored by running *Bam* HI-digested DNA on 0.9% agarose gels and using a 1.9 kb *Sal*I fragment excised from λ2842 as a probe (Fig. 20.) The variant frequency of 0.44 is in good agreement with the figure of 0.43 quoted by Aquadro *et al.* (1988).

The two variant *Sal*I sites are adjacent to one another and therefore cannot be scored independently. Lines digested with this enzyme were probed with intact λ2837 in an attempt to score both sites simultaneously, but the

Figure 19

*Cfo*I-digested DNA probed with pRA-G3, revealing a third restriction-site polymorphism (C-167) with this enzyme (see 3.2.3).

Opposite page 53

Fig. 19

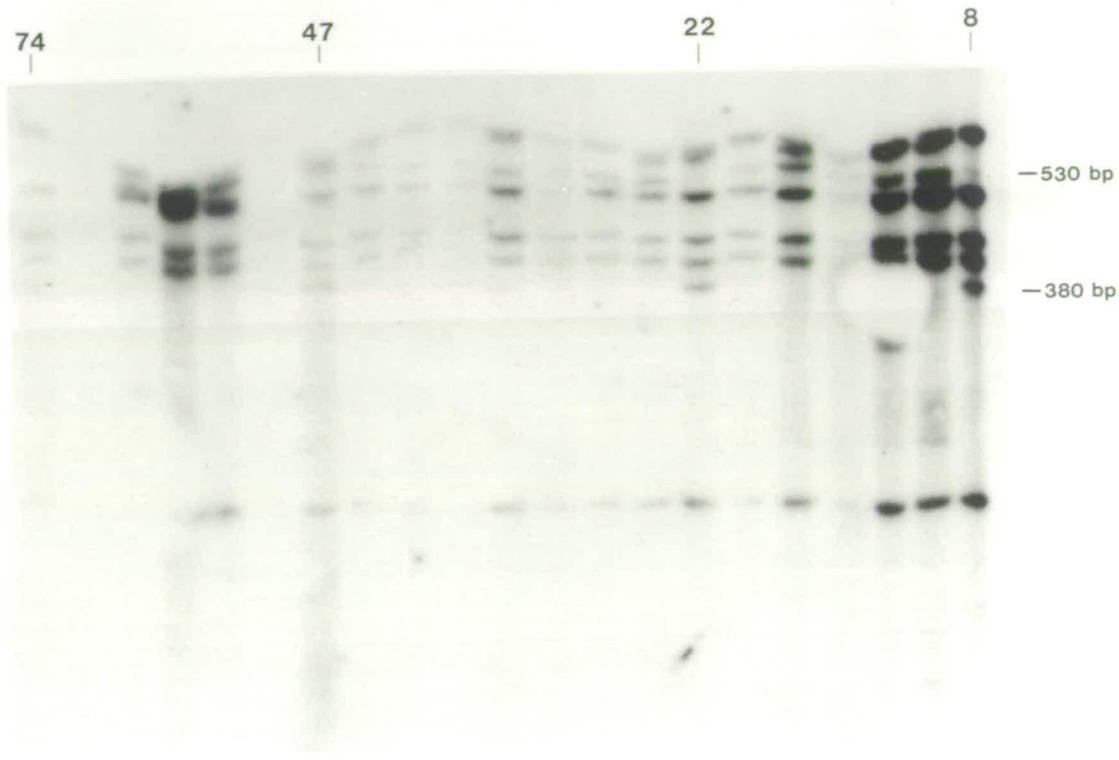
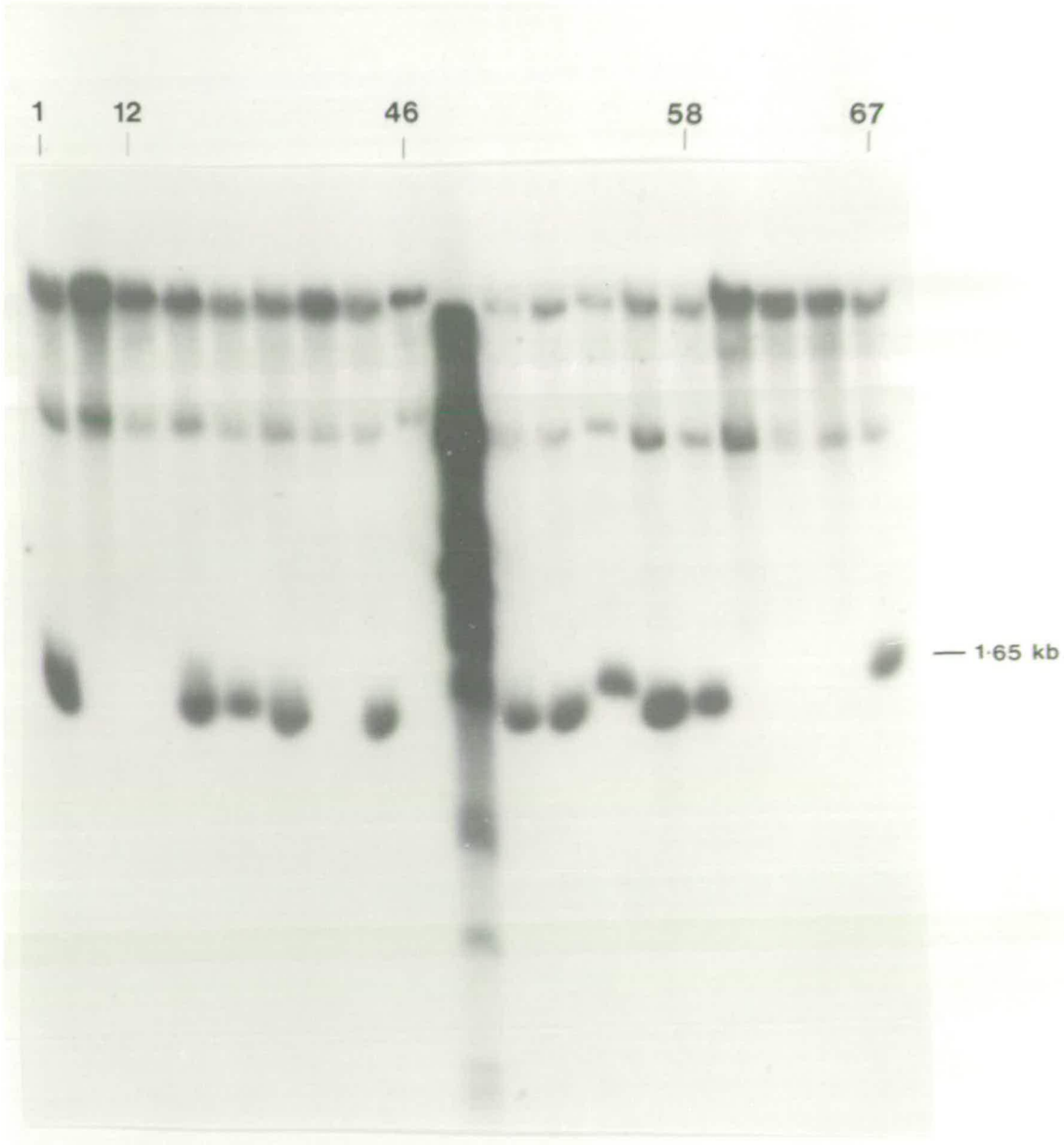


Figure 20

DNA digested with *Bam*HI, separated in a gel of 0.9% agarose and probed with 42-S3 (see 3.2.4) showing the B-178 polymorphism. Although the 17.6 kb fragment is cleaved to produce the 1.65 kb daughter fragment, its consequent reduction in size is not sufficient for resolution in this concentration of gel. This polymorphism had been described previously by Aquadro *et al.* (1988).

Opposite page 54

Fig. 20



excessive information generated by the combination of two variables compounded by the incomplete digestion of many of the lines, ultimately rendered these two sites effectively unscorable (Fig. 21).

Table A2 in Appendix IV lists the sizes and frequencies of variable and constant bands for each of these polymorphisms in the *rosy* gene region.

3.2.5. Extension across the region

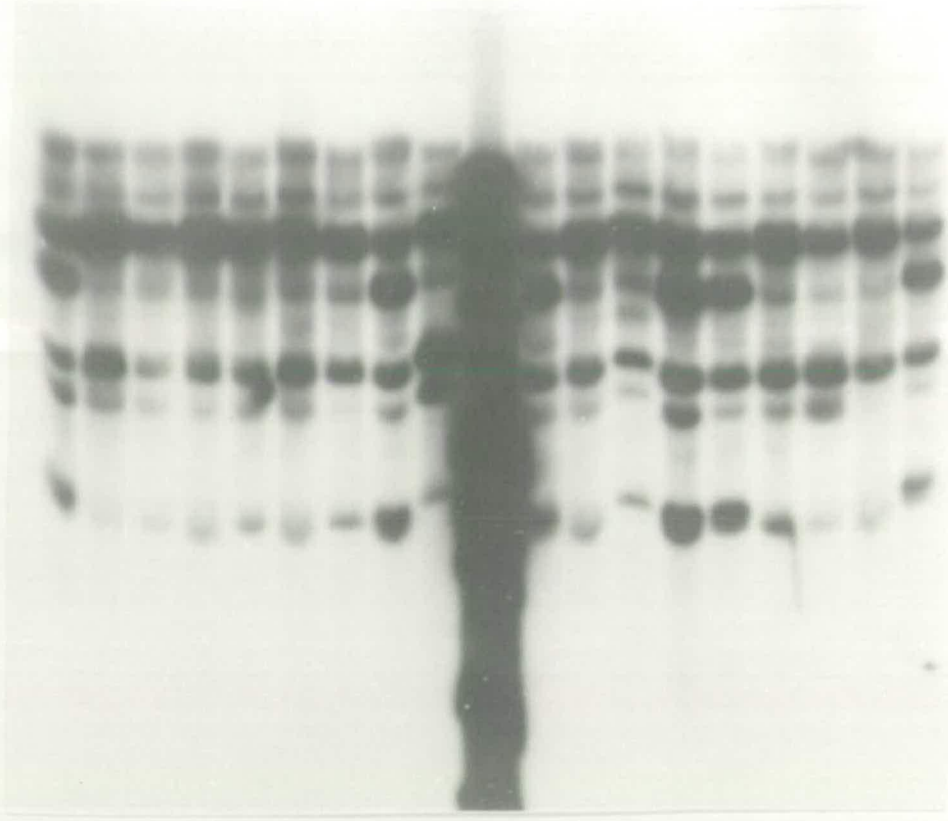
It had been an objective of this survey to repeat the analysis across the *ry-Ace* region by hybridizing the same set of membranes with a succession of different cloned probes, a procedure used to good effect in the *ac-sc* survey. This would have allowed for a more rigorous correlation of molecular distance with genetic linkage, since fragments corresponding to several mapped genes were subcloned. However, problems encountered in the removal of small fragments from the membranes precluded this aim. Use of conventional stripping treatments with sodium hydroxide resulted in a significant reduction of signal upon re-hybridization, while the alternative protocol using methanamide solution was ineffective at removing the original probe (Fig. 22.) As a result, no variants were detected near the *pic* and *Ace* gene sequences and the analysis must therefore be confined to a shorter stretch of the DNA than had been intended. This is to some extent compensated by the opportunity for resolution of extremely close sites within the *rosy* gene presented by the available DNA sequence information: for example, H-167 and A-167 are a mere 137 bp apart. In the event, the pattern of disequilibrium characteristic of the *ry-Ace* region did not seem to be obscured by the lack of comparisons covering large distances, as will be apparent from the forthcoming analysis.

Figure 21

*Sal*I-digested DNA probed with λ2837, to investigate two polymorphisms described by Aquadro *et al.* (1988). These represent successive recognition sites for this enzyme on the DNA strand (−163.2 and −158.3 on the map) which complicates their expected banding patterns; in these lines the DNA did not digest sufficiently well for the unambiguous assignment of the four allele categories to be made.

Opposite page 56

Fig. 21



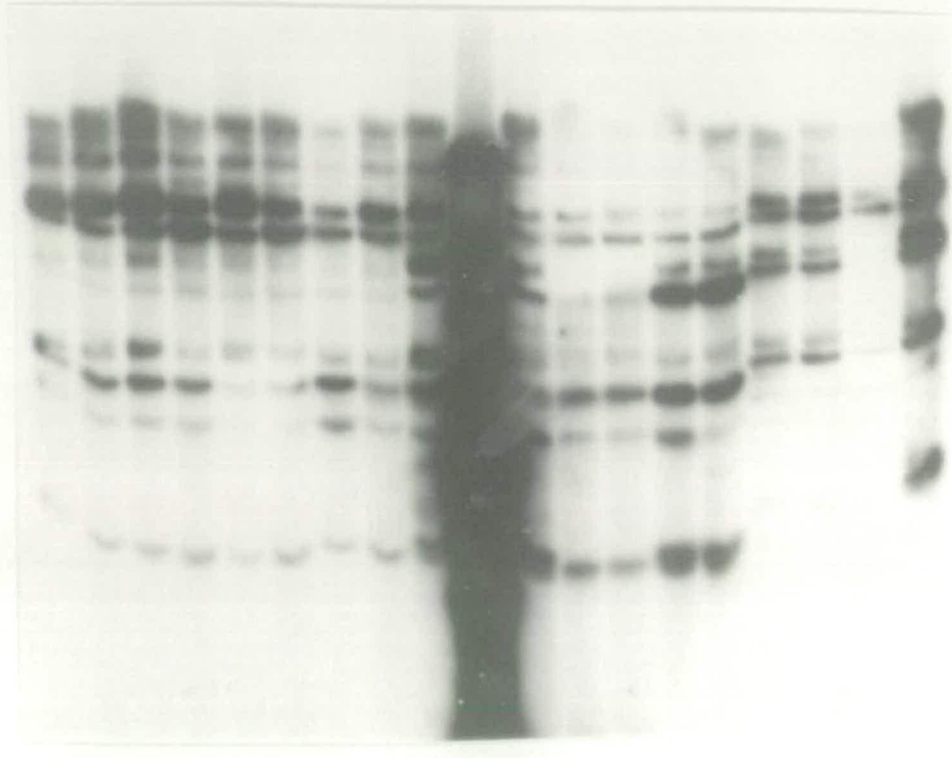


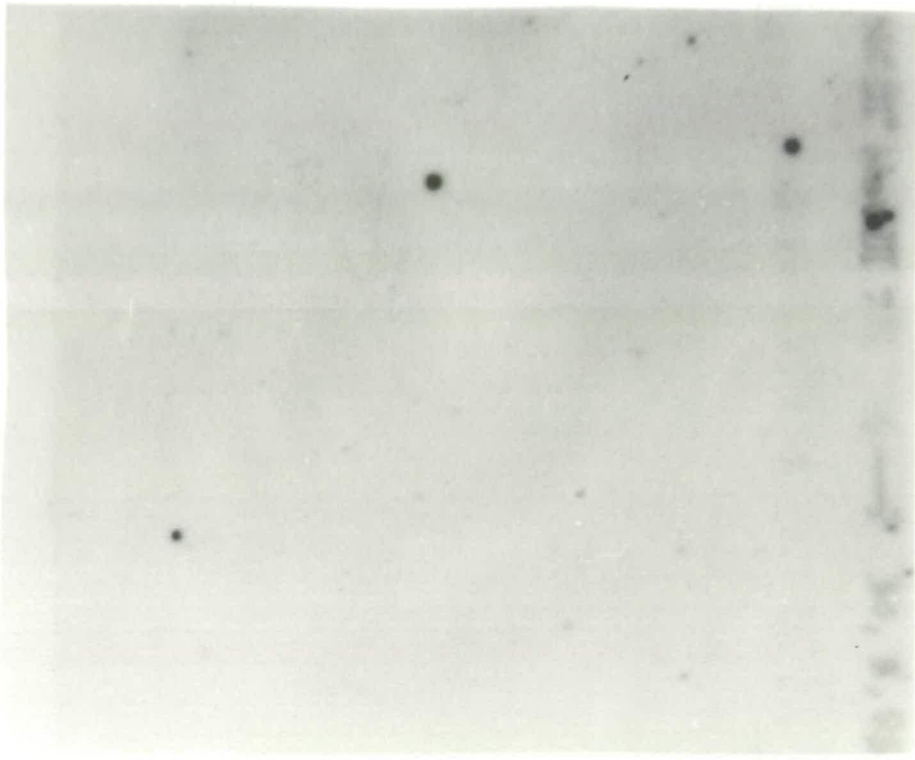
Figure 22

Comparison of the efficacy of two alternative procedures for removal of probe DNA from membranes after use, as applied to two membranes which had experienced the same hybridization conditions. Treatment with 0.4 M NaOH at 42 °C for 20 min followed by neutralization is highly effective at stripping away probe (top), but may be too harsh for some purposes as small fragments of bound genomic DNA may also be detached. The milder stripping conditions utilizing 96% methanamide, however, cannot be relied upon to remove all traces of the probe, as the overnight exposure of a treated membrane clearly demonstrates (bottom).

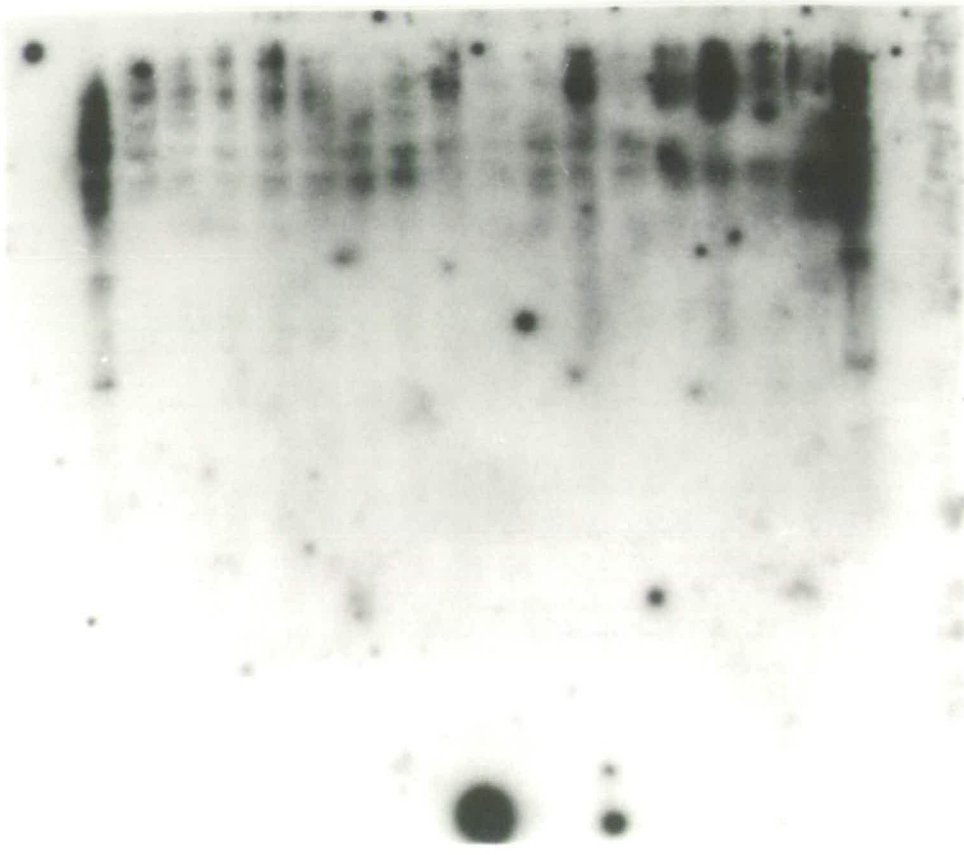
Opposite page 57

Fig. 22

NaOH



Methanamide



4.1. Measures of linkage disequilibrium

4.1.1. Two loci

The extent of non-random association between alleles at two loci may be quantified in a number of ways (Hedrick 1985, 1987, 1988; Hedrick *et al.*, 1978) other than the primary coefficient *D*, which has several undesirable characteristics. Its value is positive for alleles in coupling disequilibrium, and negative for alleles in repulsion disequilibrium; it follows that the sign of linkage disequilibrium is entirely dependent upon the system used for naming the alleles, unless a convention is adopted such as that of Langley *et al.* (1974) whereby the rarer allele is in every case assigned the lower number. The maximum possible value of *D* is ± 0.25 when the allele frequencies are 0.5 at both loci; for every other value of *p* and *q*, D_{max} is lower than 0.25. This maximum value of *D* declines not only with lower allele frequencies but also as a function of greater discrepancy between the allele frequencies, one effect of this being that much higher sample sizes are required to reject a null hypothesis of D=0 under these circumstances (Brown, 1975).

To overcome the extreme dependence of *D* on allele frequency, Lewontin (1964) suggested using the statistic

$$D' = D/D_{max}$$

as a standardized measure of disequilibrium. D_{max} may be calculated as the lesser of p_1q_2 and p_2q_1 when *D* is positive, and the lesser of p_1q_1 and p_2q_2 when *D* is negative. The maximum value of *D'* is then +1 for positive *D* and −1 for negative *D*, and is independent of allele frequency. Although *D'* can be a very useful statistic for assessing the true extent of disequilibrium for any given allele frequency, great care needs to be taken in its interpretation as its value can be artificially inflated when dealing with variants at low frequency in small samples; there is no standard statistical test for *D'* as its sampling distribution is unknown. Several authors prefer to describe observed associations in terms of *D'* while evaluating their statistical probability as

chance events by means of Fisher's exact test of independence (Langley and Aquadro, 1987; Langley *et al.*, 1988; Schaeffer *et al.*, 1988; Lado *et al.*, 1988).

In this analysis, however, I have placed most emphasis on another standardized measure of disequilibrium, following Hill and Robertson (1968). The squared correlation coefficient, calculated as

$$r^2 = D^2/p_1p_2q_1q_2$$

transforms coupling and repulsion disequilibrium alike to a positive value; although moderately sensitive to allele frequency (Hedrick, 1988), its range from zero in the case of competely random association to unity indicating maximum disequilibrium between alleles equal in frequency would seem to provide the most easily interpretable and comparable quantification of observed associations. Both r^2 and r itself have been widely used by previous experimenters (*e.g.* Charlesworth and Charlesworth, 1973; Birley, 1974; Chakravarti *et al.*, 1984; Cross and Birley, 1986; Epperson and Allard, 1987; Miyashita and Langley, 1988; Smit-McBride *et al.*, 1988; Eanes, Labate and Ajioka 1989; Eanes *et al.*, 1989). The sampling distribution of r^2 is well established: the statistic $Q=nr^2$ where n is the sample size, is distributed as χ^2 with one degree of freedom, provided that $4Nc \gg 1$ (Golding, 1984). As the expected value of r^2 is a function only of c, N_e and n, observed values of r^2 can also be used for estimating the effective population size N_e (Hill, 1981).

4.1.2. Three loci

The terms for calculating linkage disequilibrium between three loci, after taking account of the two-way associations, have been derived by Bennett (1954). By analogy with the two-locus case, three-way disequilibrium is the departure of haplotype frequency from the product of the component allele frequencies over and above that described by the three pairwise coefficients. That is, if ρ_1 and ρ_2 are the frequencies of two alleles at locus C,

$$f(A_1B_1C_1)=p_1q_1\rho_1+p_1D_{BC}+q_1D_{AC}+\rho_1D_{AB}+D_{ABC}$$
$$f(A_1B_2C_2)=p_1q_2\rho_2+p_1D_{BC}-q_2D_{AC}-\rho_2D_{AB}+D_{ABC}$$
$$f(A_2B_1C_2)=p_2q_1\rho_2-p_2D_{BC}+q_1D_{AC}-\rho_2D_{AB}+D_{ABC}$$
$$f(A_2B_2C_1)=p_2q_2\rho_1-p_2D_{BC}-q_2D_{AC}+\rho_1D_{AB}+D_{ABC}$$
$$f(A_1B_1C_2)=p_1q_1\rho_2-p_1D_{BC}-q_1D_{AC}+\rho_2D_{AB}-D_{ABC}$$
$$f(A_1B_2C_1)=p_1q_2\rho_1-p_1D_{BC}+q_2D_{AC}-\rho_1D_{AB}-D_{ABC}$$

$$f(A_2B_1C_1)=p_2q_1\rho_1+p_2D_{BC}-q_1D_{AC}-\rho_1D_{AB}-D_{ABC}$$
$$f(A_2B_2C_2)=p_2q_2\rho_2+p_2D_{BC}+q_2D_{AC}+\rho_2D_{AB}-D_{ABC}$$

from which the three-way coefficient may be computed:

$$D_{ABC}=f(A_1B_1C_1)-p_1q_1\rho_1-p_1D_{BC}-q_1D_{AC}-\rho_1D_{AB}$$

An analogue of D' can be used to relate D_{ABC} not only to its maximum but also to its minimum value, which for some allele frequencies may not be zero (Thomson and Baur, 1984):

$$D''_{ABC} = D_{ABC}-D_{min}/D_{max}-D_{min}$$

The analogous squared correlation coefficient is then

$$r^2_{ABC} = D^2_{ABC}/p_1p_2q_1q_2\rho_1\rho_2$$

and Q_{ABC} will follow a chi-square distribution, so statistical significance of three-way associations can again be expressed as χ^2 with one degree of freedom. Values of D_{ABC} for any given locus triplet are likely to be lower than pairwise disequilibria for the same loci, unless selection is stronger at the three-locus level (Hastings, 1986); as a consequence, statistically significant higher-order disequilibrium tends to be less commonly observed (Brown, 1975).

4.2. Estimation of nucleotide variability

Several methods have been detailed for the estimation of DNA sequence variability from restriction enzyme data; that of Ewens *et al.* (1981) is among the most straightforward and widely-used. If k out of m cleavage sites are not monomorphic in the sample (*i.e.* absent in at least one line) for an enzyme recognizing a sequence of j base-pairs, then the proportion of nucleotides estimated to be polymorphic is given by

$$P = k/2mj$$

The combined data from each set of restriction enzymes with equivalent j may

be substituted in this formula. The present survey used only four- and six-cutter enzymes including one six-cutter, *Ban*II, for which two positions in a cleavage site may be occupied by either of two bases, and which therefore behaves as a five-cutter. Hence the estimator becomes:

$$P = (k_4+k_5+k_6)/(8m_4+10m_5+12m_6)$$

From this figure can be derived in turn an estimate of the proportion of any two randomly-chosen homologous DNA tracts which will possess different nucleotides at any one position, *i.e.* the estimated average per nucleotide heterozygosity, as

$$\theta = P/\ln n$$

where *n* is the sample size, or number of DNA segments surveyed. The above formulae were derived under the assumption of selective neutrality according to a Fisher-Wright model, whereupon $\theta=4N_e\mu$ where μ is the mutation rate to selectively neutral alleles. This treatment yields similar estimates in practice to other published methodologies (Engels, 1981; Hudson, 1982). An alternative measure of variability is provided by the coefficient of nucleotide diversity, given by

$$\pi = (n/n-1)[(-\ln S)/j]$$

where *S* is the estimated proportion of shared restriction sites (Nei and Tajima, 1981). The two estimators may yield different values when the distribution of allele frequencies for variable sites is skewed, since θ is a function only of the number of segregating sites and takes no account of their allele frequency.

4.3. The *ac-sc* region

The allelic state of each polymorphic site scored in the sample of 44 X chromosomes is given as presence or absence of a restriction site or insertion (Table 1). Only sites where the rarer allele was present more than twice in the sample are tabulated; variant frequencies are presented for each site at the foot of the table. Blank entries indicate lines which were not investigated for that particular site, while lines investigated but not scored are

Table 1

Haplotypes of all lines scored for polymorphic events in the *ac-sc* complex, recorded as presence (+) or absence (-) of a restriction site or insertion. A blank entry indicates that the line was not investigated for that site, while sites that could not be scored in a particular line are entered as *n*. The nine identified haplotypes are listed in the right-hand column; those marked with hats (^) were inferred from the data.

Opposite page 62

Table 1

achaete-scute variation

Site	*Cfo* I	*Msp* I	*Bgl* II	*Hae* III	*Xba* I	*Insertion*	*Bgl* II	
Position	67	55	48	28	11	IV	-19	
Line								Haplotype
NC1	+	+	-	+	+	-	+	1
3	+	+	-	+	+	-	-	2
4	+	+	-	+	-	-	+	3
8	-	-	+	-	-	+	-	4
11	+	+	-	+	-	-	-	5
12	-	-	+	-	-	+	-	4
14	+	+	-	+	+	-	+	1
15	-	-	+	-	-	+	-	4
17	+	+	-	+	-	-	-	5
18	+	+	-	+	+	-	+	1
19	+	+	-	+	+	-	+	1
21	+	+	+	+	-	-	-	6
24	+	+	-	+	+	-	+	1
25	+	+	-	+	+	-	+	1
26	-	-	+	-	-	+	-	4
27	+	+	+	+	-	n	n	6^
28	-	+	+	+	-	-	-	7
31	+	+	-	+	-	-	+	3
33	+	+	-	+	+	-	+	1
34	-	-	+	-	-	+	-	4
35	n	n	+	+	-	-	-	6/7^
36	+	+	-	+	-	-	+	3
37	-	+	+	+	-	-	-	7
38	+	+	-	+	-	+	-	8
41			-		+	-	+	1^
47	+	+	-	+	+	-	-	2
49	+	+	-	+	+	-	-	2
50	n	n	-	n	-	-	+	3^
51	n	n	-	+	-	-	-	5^
54	+	+	-	+	+	-	-	2
57	-	+	+	+	-	-	-	7
60	n	+	+	+	-	-	-	6/7^
76	n	n	+	+	-	-	-	6/7^
92	+	n	-	+	-	-	-	5^
93	+	+	-	+	-	-	+	3
94	-	-	+	-	-	n	n	4^
100		+	-	+	-	-	-	5^
102		+	-	+	-	-	+	3^
103	+	+	+	+	-	-	-	6
104	+	+	-	+	+	-	+	1
127	+	+	+	+	-	-	-	6
138		+	-	+	+	+	+	9
152	+	+	-	+	-	-	+	3
158		+	-	+	+	-	+	1^
Freq.	0.26	0.16	0.36	0.14	0.34	0.17	0.43	

denoted by *n*. Haplotypes are numbered in order of first occurrence proceeding down the table; the numbering of the 44 NC lines is historical, not sequential as in published work (see Appendix III).

4.3.1. Nucleotide variability

Table A3 in Appendix IV summarizes the results of the *ac-sc* survey with all the restriction enzymes used. However, only polymorphisms detected for the first time in these experiments were used in the estimation of nucleotide variability, thereby avoiding bias from the selection of known polymorphisms. Using the formulae given above for j=4 and 5 only,

$$P = 5/[(8\times143)+(10\times13)]$$
$$= 0.003\ 9$$

and

$$\theta = 0.0039/\ln 16 \qquad \text{(Mean sample size = 16)}$$
$$= 0.001\ 4$$

The variance of θ, calculated from equations (19) and (23) of Hudson (1982) gives a standard error of 6.0×10^{-3} or 1.6×10^{-4}, assuming no recombination or free recombination respectively. These estimates are independent of those arrived at by Beech and Leigh Brown (1989) of P = 0.007 4 and θ = 0.002 4. Nucleotide diversity is estimated at π = 0.002 4 from these data, compared to 0.003 3 from those of Beech and Leigh Brown (1989).

4.3.2. Nonrandom associations

Table 2 shows the linkage disequilibrium for every pairwise combination of the variant sites discovered in the *ac-sc* region, arranged in increasing order of their separation in kilobases. The estimators D, D' and r^2 were used as measures of disequilibrium, the last of these multiplied by the sample size n (see 4.1.1) to give Q which was then used to determine the probability p(1 d.f.) of the observed associations being due to chance. The data are notable for the large number of highly significant r^2 values, even at the greater distances; this may be further illustrated by a graph of r^2 against distance (Fig. 23). All data points shown as circles represent statistically significant disequilibria. The linear regression of r^2 on distance has gradient $-0.000\ 41\ \text{kb}^{-1}$, not significantly different from zero. The clustering of some data points suggests a certain degree of autocorrelation due to the

Table 2

Coefficients of linkage disequilibrium for all possible pairwise comparisons of variant sites described in the *ac–sc* region, arranged in increasing order of distance between the sites. For a description of D', r^2 and Q see 4.1.1 in text.

Opposite page 64

Table 2

Disequilibrium at the *ac-sc* complex

Comparison	Distance/kb	D	D'	r^2	Q	p (1d.f.)	n
G-19 x IV	5	-0.047 6	-0.67	0.066 7	2.8	>0.075	42
G48 x M55	7	-0.099 7	-1	0.321	12	<0.001 **	38
C67 x M55	12	0.132	1	0.593	20	<0.000 5 **	33
X11 x H28	17	0.047 6	1	0.083 3	3.5	>0.05	42
G48 x C67	19	-0.163	-1	0.582	20	<0.000 5 **	34
G48 x H28	20	-0.088 4	-1	0.271	11	<0.001 **	42
X11 x IV	25	-0.035 7	-0.60	0.040 0	1.7	>0.15	42
M55 x H28	27	0.133	1	1	38	<0.000 5 **	38
X11 x G-19	30	0.109	0.53	0.211	8.8	<0.005 **	42
X11 x G48	37	-0.124	-1	0.296	13	<0.000 5 **	44
H28 x C67	39	0.130	1	0.595	20	<0.000 5 **	34
H28 x IV	42	-0.103	-1	0.673	27	<0.000 5 **	40
X11 x M55	44	0.058 2	1	0.109	4.2	<0.05 *	38
H28 x G-19	47	0.05	1	0.095 2	3.8	>0.05	40
X11 x C67	56	0.093 4	1	0.196	6.7	<0.01 **	34
G48 x IV	62	0.063 5	0.57	0.131	5.5	<0.025 *	42
G48 x G-19	67	-0.143	-1	0.375	16	<0.000 5 **	42
M55 x IV	69	-0.112	-1	0.668	24	<0.000 5 **	36
M55 x G-19	74	0.061 7	1	0.129	4.6	<0.05 *	36
C67 x IV	81	-0.109	-0.78	0.419	13	<0.000 5 **	32
C67 x G-19	86	0.102	1	0.228	7.3	<0.01 **	32

17/21 comparisons significant *. 14/21 highly significant **.

Figure 23

Graph of absolute linkage disequilibrium, measured as r^2, against physical distance separating variant sites in the *ac-sc* region (data from Table 2). All statistically significant disequilibria ($p<0.05$) are plotted as circles, the remainder as triangles. The regression coefficient has variance 5.7×10^{-6} kb^{-2}.

Opposite page 65

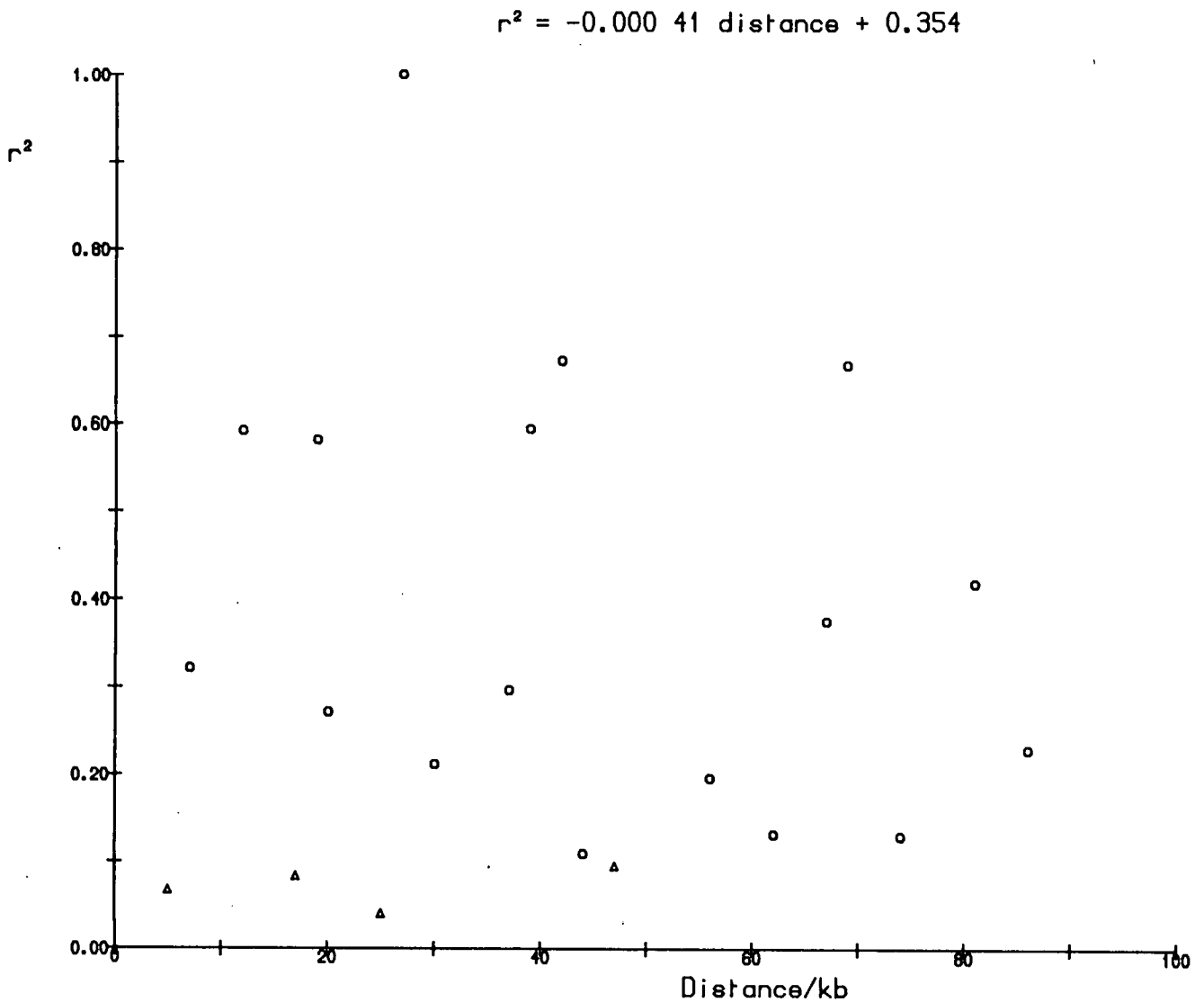

Regression of disequilibrium on distance in the *ac-sc* region
r^2 = -0.000 41 distance + 0.354
r^2
1.00
0.80
0.60
0.40
0.20
0.00
0
20
40
60
80
100
Distance/kb

Fig. 23

nonindependence of the pairwise comparisons, but this is not too extreme.

A more important aspect of the data is the large proportion of unit D' values, indicating the absence of one gametic class from the sample and consequent maximal disequilibrium for the respective allele frequencies; clearly, disequilibrium is a property of the entire complex with many alleles segregating in discrete combinations.

4.3.3. Three-way associations

The interesting results at the two-locus level make desirable a closer scrutiny of the data for possible higher-order interactions. Three-way disequilibria were computed using a program kindly made available by Dr B.S. Weir, and values of D_{ABC} and Q_{ABC} are shown in Table 3 (not ordered by distance) for sample sizes of n. The values of r^2_{ABC} would be arrived at by dividing Q/n. Although not as striking as the two-locus disequilibria, several significant associations are revealed by this analysis and once more show no apparent relationship to distance. All include variants involved in strong pairwise disequilibria, but their importance is nevertheless undiminished: as alluded to earlier (4.1.2), the magnitude and significance of three-way D values depends only on the levels of association over and above that due to the pairwise values. As a result, there are very few published instances of significant three-locus associations to date.

4.3.4. Haplotype diversity

Table 1 shows that there are at least nine different haplotypes present in the sample, numbered in sequence down the table. Even if we consider only the complete 7-locus haplotypes it is clear that only a small fraction of all possible combinations is represented and their distribution is markedly skewed. In 12 of the 44 lines the complete haplotype is unknown, but from the independent evidence of linkage disequilibrium extending across the entire set of sites (Table 2) with many unit D' values, coupled with the observed frequency distribution of the complete haplotypes (Table 1) it would be a plausible assumption that only these nine haplotypes are in fact present in the sample. Extending this argument, one could infer the most parsimonious 7-locus haplotypes in all the lines that are incompletely scored; for example, line 158 is considered more likely to be C67(+) (haplotype 1) than

Table 3

Coefficients of three-way linkage disequilibrium, D_{ABC} between all triplet combinations of variant sites in the *ac-sc* region. D_{ABC} may be calculated as described in 4.1.2; the program for the calculation was provided by Dr B.S. Weir. The statistic Q again represents χ^2 with one degree of freedom. The sites are not arranged by distance in the table, but the maximum distance between sites is given for those combinations that are significant.

Opposite page 67

Table 3

Three-way linkage disequilibrium

Comparison			D	Q	n	p (1 d.f.)		Maximum distance/kb
X11	G-19	G48	-0.0147	0.96	42			
X11	G-19	H28	0.0037	0.33	40			
X11	G-19	M55	0.0094	1.01	36			
X11	G-19	IV	-0.0125	1.50	42			
X11	G-19	C67	0.0137	1.06	32			
X11	G48	H28	-0.0113	1.12	42			
X11	G48	M55	-0.0153	1.30	38			
X11	G48	IV	0.0108	0.86	42			
X11	G48	C67	-0.0220	1.58	34			
X11	H28	M55	0.0398	4.46	38	<0.05	*	44
X11	H28	IV	-0.0316	3.63	40			
X11	H28	C67	0.0293	2.93	34			
X11	M55	IV	-0.0369	4.05	36	<0.05	*	69
X11	M55	C67	0.0301	2.87	33			
X11	IV	C67	-0.0234	1.74	32			
G-19	G48	H28	-0.0150	1.31	40			
G-19	G48	M55	-0.0206	1.56	36			
G-19	G48	IV	0.0113	0.74	42			
G-19	G48	C67	-0.0317	2.27	32			
G-19	H28	M55	0.0446	4.46	36	<0.05	*	74
G-19	H28	IV	-0.0356	4.10	40	<0.05	*	47
G-19	H28	C67	0.0317	2.69	32			
G-19	M55	IV	-0.0416	4.48	36	<0.05	*	74
G-19	M55	C67	0.0327	2.63	31			
G-19	IV	C67	-0.0254	1.79	32			
G48	H28	M55	-0.0682	7.21	38	<0.01	**	27
G48	H28	IV	0.0591	7.58	40	<0.01	**	62
G48	H28	C67	-0.0513	3.50	34			
G48	M55	IV	0.0643	7.35	36	<0.01	**	69
G48	M55	C67	-0.0501	3.27	33			
G48	IV	C67	0.0488	3.33	32			
H28	M55	IV	-0.0808	8.53	36	<0.005	**	69
H28	M55	C67	0.0841	6.93	33	<0.01	**	39
H28	IV	C67	-0.0781	7.05	32	<0.01	**	81
M55	IV	C67	-0.0785	6.78	31	<0.01	**	81

C67(-) which would create a previously unrecorded 10th haplotype. Inferred haplotypes are denoted with hats (^) in Table 1.

Fig. 24 shows the probable distribution of haplotypes in the *ac-sc* region. The most common haplotype is 1, and most of the others display close homology with this species; the one notable exception is 4, which is at high frequency but is the exact complement of 1. These observations may be made notwithstanding the uncertainty over haplotypes in 12 lines.

The relatedness of the haplotypes is depicted by the unrooted tree in Fig. 25, from which a number of possible phylogenies could be deduced. Most of the differences between haplotypes can be accounted for by single mutational steps; it is unlikely that recombination has played a major rôle in the evolution of this haplotype diversity. The distribution of insertion IV, however, is anomalous: its disjunct endemism, or appearance on three otherwise unrelated haplotypes, may indicate a multiple event such as recurrent transposition to this site. Another curious feature is why haplotype 4 has been entirely conserved while its complement, 1, is accompanied by a veritable 'quasispecies' or family of haplotypes related to it in simple fashion.

4.3.5. Insertion-deletion variation

The question of insertion-deletion variation in this region has been dealt with at length by Beech and Leigh Brown (1989). Although insertion IV from their survey was incorporated into this work, no new high-frequency sequence-length variation was uncovered here, only two unique variants being attributable to insertions. A lower frequency of large insertions relative to nucleotide polymorphisms, especially on the X chromosome, is broadly what one would expect if the former are slightly deleterious.

4.4. The /S12-*ry* region

Alleles for each of the five polymorphic restriction sites in and around the *rosy* gene, and for the polymorphic *Bam*HI site of Aquadro *et al.* (1988) beyond the /S12 transcript, are shown for the sample of 72 lines in Table 4. Sequence data from the GENBANK database allowed the molecular localization of the four-cutter polymorphic sites down to individual base changes with reasonable certainty; the ordinates of the sites in Fig. 7 and

Figure 24

Genotypes of the nine multilocus arrays identified for the seven variant sites described in the *ac-sc* region in the sample of 44 X chromosomes surveyed. The lines possessing a particular haplotype have been arranged alongside it to indicate the frequency distribution of the haplotypes; although twelve of these have been inferred from the data, the distribution would not differ greatly if these were omitted.

Opposite page 69

Haplotype no.	Genotype	Lines
1	++-++-+	001, 014, 018, 019, 024, 025, 033, 041, 104, 158
2	++-++--	003, 047, 049, 054
3	++-+--+	004, 031, 036, 093, 102, 152
4	--+--+-	008, 012, 015, 026, 034, 094
5	++-+---	011, 017, 051, 092, 100
6	+++----	021, 027, 103, 127, ???
7	-+++---	028, 057, 037, ???, ???
8	++-+-+-	038
9	n+-++++	138

Fig. 24

Figure 25

Unrooted phylogeny for the nine haplotypes identified for seven polymorphic sites across the *ac-sc* complex, by the criterion of the minimum number of mutational events between near-neighbours; length of the lines is proportional to the number of differences between adjoining haplotypes. The requisite mutational changes are shown alongside the lines; as shown by the dotted lines, the transition between 1 and 5 could have occurred by either of two routes with equal probability. The simple homology of all arrays other than 4 has made it possible to construct the tree without incorporating recombination; however, it is then necessary to postulate a multiple origin for insertion IV.

Opposite page 70

Fig. 25

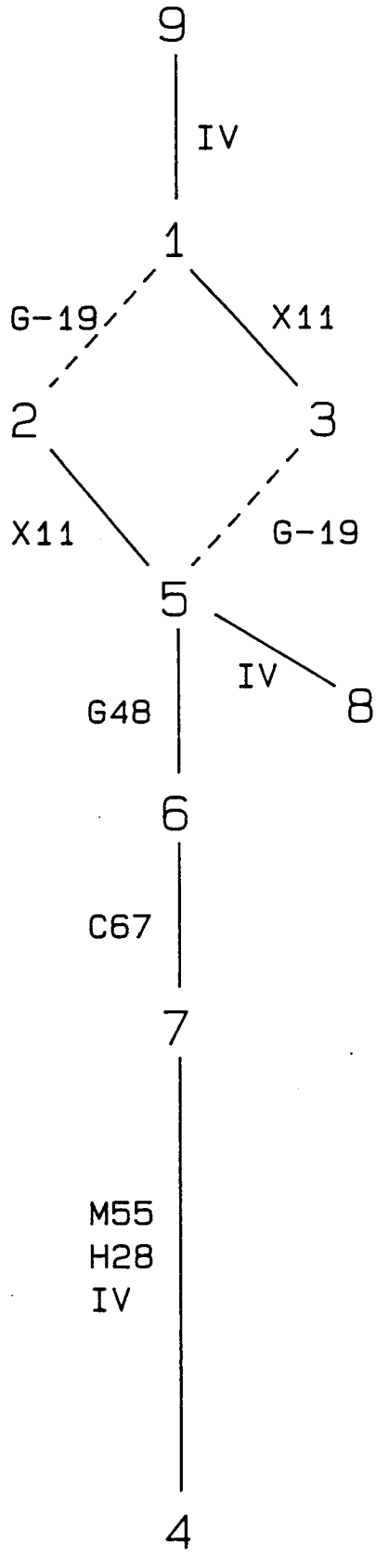


Table 4

Haplotypes of all lines investigated for the seven restriction-site variants described in the *rosy* gene region. Sites denoted by *n* were not scorable; a gap is left where a line was not screened for a given site. The large proportion of unscored lines for the *Alu*Ib variant is due to its position just 15 bp distal to the other polymorphic *Alu*Ia site and coinciding with the limit of the pRA-G3 probe, so that the allelic state of this variant only becomes apparent when the inner A-167a site is absent (-) in that line.

Opposite page 71

Table 4

Restriction-site variation in the JS12-*ry* region

Site	*Bam* HI	*Cfo* I	*Cfo* I	*Cfo* I	*Hae* III	*Alu* Ia	*Alu* Ib
Position	-178.2	-169.3	-168.3	-167.3	-167.0	-166.9	-166.9
Line							
NC 1	+	+	-	-	-	-	+
2	+	-	-	-	n	n	n
3	+	-	-	-	+	-	-
6	-	-	-	-	+	n	n
8	-	-	-	+	+	-	-
9	+	-	-	-	+	+	n
10	+	-	-	-	+	+	n
11	-	-	-	-	+	+	n
12	-	-	-	-	+	n	n
14	+	-	-	-	-	-	+
16	-	-	-	-	-	n	n
19	+	-	-	-	-	-	+
22	+	-	-	+	+	-	-
23	+	-	-	-	+	+	n
25	+	-	-	-	-	n	n
34	-	-	+	-	+	+	n
35	+	-	-	-	+	+	n
36	+				-	n	n
38	+				+	+	n
41	-	-	-	-	-	+	n
46	-	-	-	-	-	-	+
47	+	-	-	+	+	-	-
49	+	-	-	-	+	+	n
50	+	-	-	-	+	+	n
51	+	-	-	-	-	n	n
54	+	-	-	-	-	-	+
56	+	-	-	-	-	-	+
58	+	+	-	-	-	-	+
61	-	-	-	-	+	+	n
62	-				+		
63	-	n	n	n	+	-	+
64	-	-	-	-	+	+	n
66	-	-	-	-	+	n	n
67	+	-	-	-	-	-	+
69	+	-	-	-	-	-	+
71	-	-	-	-	-	+	n
72	-	-	+	-	-	-	+
74	+	-	-	+	+	-	-
76	+	-	-	-	-	-	+
77	+	-	-	-	-	+	n
79	-	-	+	-	-	-	+
80	+	-	-	-	+	+	n
88	-				+	+	n
89	+	-	-	-	-	-	+
94	-	n	n	n	+	-	+
95	-				+	+	n
99	+	-	-	-	-	+	n
100	-	-	-	-	+	+	n
103	+	-	-	-	+	+	n
104	-	-	-	-	-	-	+
114	-	-	-	-	-	-	+
118	-	-	-	-	-	+	n
120	+	-	-	-	+	+	n
122	+				+	+	n
126	-				+	n	n
132	+	-	-	-	-	-	+
133	+	-	-	-	+	-	+
134	-	-	+	-	-	+	n
135	+	-	-	-	+	+	n
136	-	-	-	-	-	+	n
137	+	-	-	-	+	+	n
139	+	-	-	-	-	-	+
140	-	-	-	-	-	+	n
141	-	-	-	-	+	+	n
145	+	-	-	-	-	-	+
150	-	-	-	-	-	+	n
151	+	-	+	-	+	+	n
152	-	-	-	-	+	+	n
153	-	-	-	-	+	+	n
156	+	-	-	-	-	-	+
157	-	-	-	-	+	+	n
158	+	-	-	n	+	+	n
Freq.	0.44	0.03	0.08	0.06	0.45	0.44	0.19

Table 4 are relative to those of Bender *et al.* (1983). No major sequence-length variation was evident in the lines sampled, although as mentioned earlier (3.2.3), some of the minor variation in band mobilities from *Cfo*I digests is suggestive of small insertions or deletions.

4.4.1. Nucleotide variability

4.4.1.1. Distribution of site polymorphism

We may assume that the polymorphic restriction sites detected here are most likely to be located at positions in the DNA strand where the requisite change in recognition sequence could occur as a single mutational step from the published sequence; on this premiss, the events underlying the *Hae*III and *Alu*I polymorphisms are thought to be a mutation from A→C at position 4 011, and mutations away from A at positions 4 150 and 4 165 of the sequence in Keith *et al.* (1987). Such changes would give rise to fragments of approximately the sizes observed (see 3.2.1, 3.2.2), placing these three polymorphisms very close together just beyond the 3′ end of the proximal *rosy* transcript.

Similarly, the three *Cfo*I polymorphic sites have been ascribed to changes from C→G at position 1 757, A→G at 2 692 and A→G at position 3 742 of Keith *et al.* (1987), all within the coding portion of the gene. The latter two are, if this assessment is correct, in the third positions of codons and are therefore silent base substitutions. The *Bam*HI polymorphic site, meanwhile, is well beyond the end of the *rosy* transcript; hence, only one out of these six variants is believed to result in an amino-acid substitution.

4.4.1.2. Heterozygosity

The data used in the calculation of heterozygosity at the *rosy* locus are presented in table A4 of Appendix IV, from which the proportion of polymorphic sites is estimated as

$$P = 9/[(8\times37)+(10\times2)]$$
$$= 0.028$$

and

$$\theta = 0.028/\ln 44 \qquad \text{(Mean sample size = 44)}$$
$$= 0.007\ 5$$

with standard error 6.2 x 10^{-3} or 3.8 x 10^{-4}, assuming no recombination or free recombination (Hudson, 1982). As for *ac-sc,* the standard error is larger for the case of no recombination where the variance of P may contribute substantially to the variance of θ. The estimate of nucleotide diversity for this region is π=0.014.

4.4.2. Nonrandom associations

The resolution of a superficially single *Cfo*I polymorphism into two distinct ones reduces the frequency of the C-169.3 variant to twice out of 65 lines, effectively too small to be useful in the estimation of disequilibrium. This site was therefore omitted from the analysis, and disequilibrium calculated for all other pairwise combinations of variants whose frequencies exceeded 0.05. A complication arises in the case of the two adjacent polymorphic *Alu*I sites: the outer of these, A-167b, coincides with one end of the probe and therefore can only be scored in lines lacking the inner (A-167a) site, since if the latter is cleaved only 15 bp of homology will exist between the junction fragment and probe, an amount not resoluble by autoradiography under the conditions used. This means that disequilibrium between the two cannot be calculated as D automatically becomes zero. The A-167b site was included in the analysis for pairwise comparison with the other sites, risking the inevitable bias if, as expected, the allelic state of the unscored lines at this position is not random.

Table 5 summarizes the pairwise linkage disequilibrium involving these restriction site polymorphisms. Although the number of comparisons is fewer than obtained for *ac-sc,* the data still show clearly the much lower values of r^2 and D' to be found in this region of the genome: five out of 14 comparisons are significant from these data alone, 5/17 if the data of Aquadro *et al.* (1988) are included, but no significant disequilibrium is found between sites separated by more than 500 bp. Moreover, the large number of lines surveyed proves that this is not a limitation of sample size. In certain cases, notably those involving the *Cfo*I variants, r^2 is predictably reduced by their low individual frequencies; but a consideration of D' values shows that in eight out of 14 comparisons, even including four of the abovementioned category all four classes of gamete are present in the sample, symptomatic of effective linkage equilibrium for most sites over 1 kb apart.

Table 5

Disequilibrium coefficients for all pairwise comparisons of variable restriction sites in the region of the /S12 and *rosy* loci in two independent surveys. The data of Aquadro *et al.* (1988) were incorporated since they were investigating the same NC population and one of the variable sites found (B-178) is common to both surveys. In the top part of the table only 14 comparisons are entered instead of the possible 15 because no disequilibrium can be ascertained between the two adjacent *Alu*I sites, for reasons explained in 4.4.2.

Opposite page 74

Table 5

Disequilibrium in the *JS12-ry* region

Comparison	Distance/kb	D	D'	r^2	Q	p(1 d.f.)	n
H-167 x A-167a	0.14	0.110	0.46	0.198	12	<0.000 5 **	62
H-167 x A-167b	0.16	-0.130	-1	0.540	15	<0.000 5 **	27
C-167 x H-167	0.27	0.033 3	1	0.072 5	4.4	<0.05 *	61
C-167 x A-167a	0.40	-0.039 7	-1	0.094 1	5.2	<0.025 *	55
C-167 x A-167b	0.42	-0.128	-1	0.762	19	<0.000 5 **	25
C-168 x C-167	1.0	-0.005 20	-1	0.006 05	0.38	>0.5	62
C-168 x H-167	1.3	-0.006 94	0.17	0.002 53	0.15	>0.6	60
C-168 x A-167a	1.4	0.004 15	0.10	0.000 855	0.048	>0.8	56
C-168 x A-167b	1.4	0.016	1	0.021 7	0.54	>0.4	25
B-178 x C-168	10	-0.030 7	-0.66	0.053 4	3.4	>0.05	63
B-178 x C-167	11	0.010 9	0.40	0.008 12	0.50	>0.4	62
B-178 x H-167	11	-0.026 7	-0.12	0.011 5	0.79	>0.3	69
B-178 x A-167a	11	-0.053 6	0.29	0.048 0	3.0	>0.075	62
B-178 x A-167b	11	-0.017 8	0.32	0.010 1	0.27	>0.6	27

Also from Aquadro *et al.*, 1988:

S-163 x S-158	5	0.031	1	0.049 2	2.9	>0.075	60
B-178 x S-158	15	0.016 6	0.092	0.005 09	0.31	>0.5	60
B-178 x S-158	20	-0.007	-0.038	0.002 18	0.13	>0.7	60

5/17 comparisons significant *, 3/17 highly significant **.

The graph of r^2 against distance and recombination fraction (Fig. 26) contrasts sharply with the analysis for *ac-sc,* this time illustrating a pronounced falloff of disequilibrium with distance. Included among the data points (as squares) are the r^2 values calculated from the data of Aquadro *et al.* (1988) for the three variants whose frequencies exceeded 0.05 from their study of the same NC population. Two of these were the *Sal*I variants on the 3' side of the *rosy* gene, so the combined sites cover a range of 20 kb spanning the *l*S12, *ry, snake* and *hsc*2 transcriptional units.

Fitting a regression to the points is not straightforward as the trend is patently non-linear. In deducing the exact form of the negative relationship a number of alternatives may be considered including an exponential, parabolic or cubic function, or a linear decline to zero. As disequilibrium is known to decay exponentially with time we might also expect it to do so with distance, since time elapsing in generations and increasing distance along a chromosome are analogous in their effect of increasing the probability of recombination events. Accordingly a graph of ln r^2 against distance was plotted, again including the data of Aquadro *et al.* (1988) plotted as squares (Fig. 27).

The points fall conspicuously around two distinct linear trends: those representing all pairwise comparisons of the four most tightly-linked sites within 1.5 kb show a very sharp decline, described by the equation

$$\ln r^2 = -3.76 \text{ distance} - 0.562$$

with a variance for the regression coefficient of 0.699 kb^{-2}, but for all comparisons involving at least one of the more distant markers (>5 kb separation) a much more gradual regression is observed, estimating

$$\ln r^2 = -0.228 \text{ distance} - 1.60$$

with variance 0.003 93 kb^{-2}. The striking dichotomy of these groupings demonstrates a strong element of autocorrelation in the disequilibrium data from the *rosy* region.

Figure 26

Graph of absolute disequilibrium (r^2) against physical separation and genetic map distance for pairs of variants described in the *l*S12, *ry, snake* and *hsc*2 region from the combined data of this study (circles) and that of Aquadro *et al.* (1988) plotted as squares. Genetic map distances were computed assuming a mapping function of 1.1×10^{-3} m.u. kb^{-1} (Chovnick *et al.,* 1976).

Opposite page 76

Fig. 26

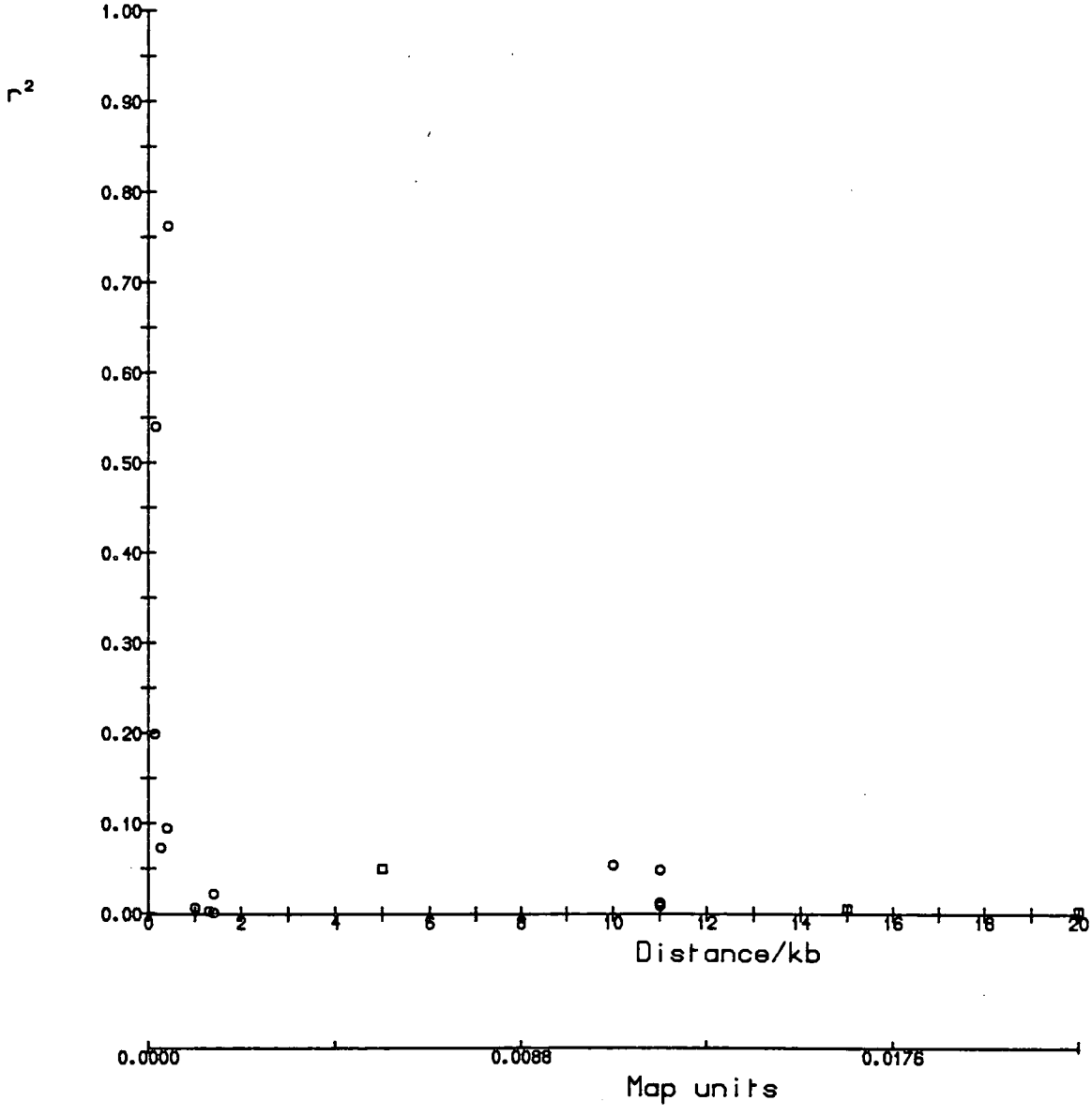
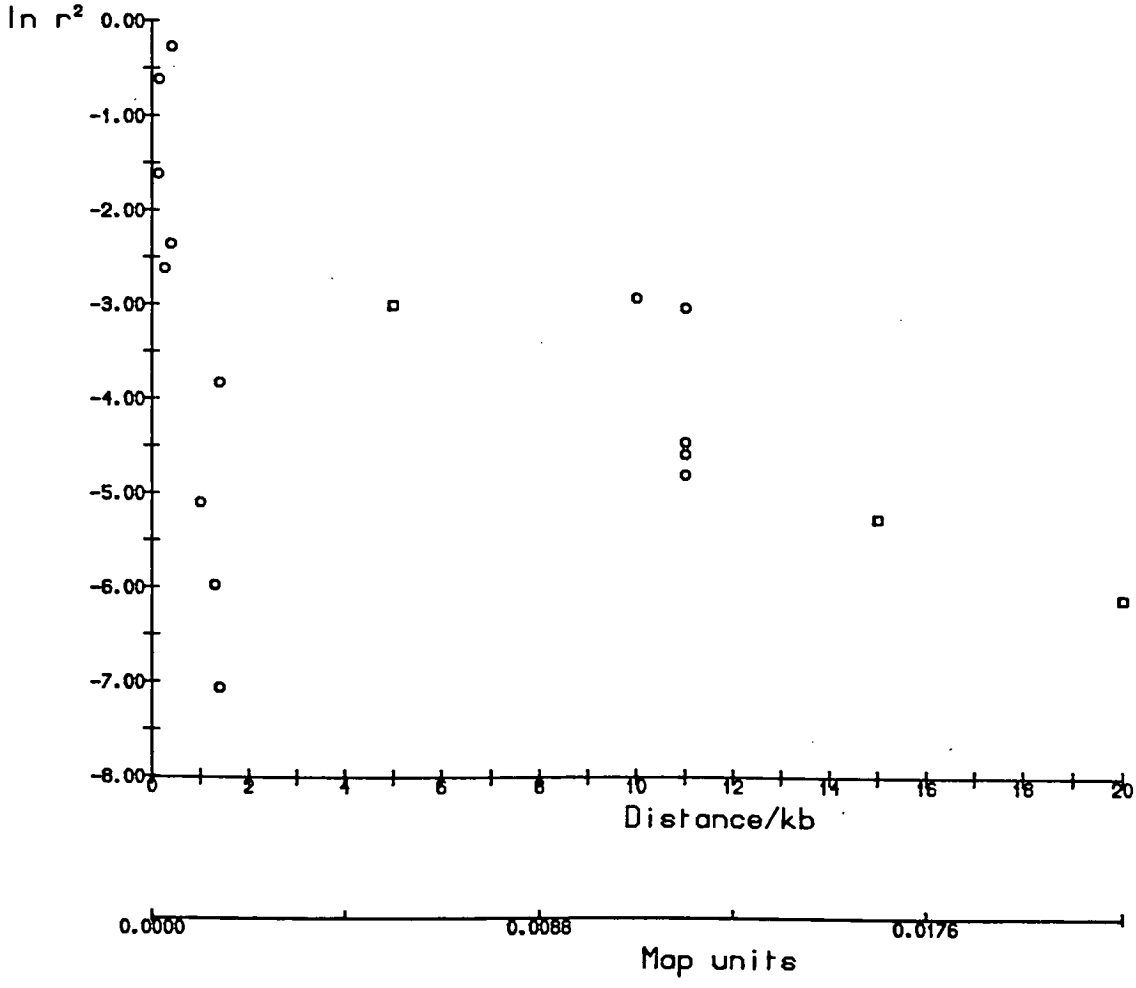

r²
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0
2
4
6
8
10
12
14
16
18
20
Distance/kb
0.0000
0.0088
0.0176
Map units

Figure 27

Graph of ln r^2 against distance in kilobases and genetic map units in the rosy regi plotted for my data (circles) and those of Aquadro *et al.* (1988).
Regression equations for the two sets of clustered data points separately are given in the text (4.4.2).

Opposite page 77

Fig. 27


ln r²
0.00
-1.00
-2.00
-3.00
-4.00
-5.00
-6.00
-7.00
-8.00
0
2
4
6
8
10
12
14
16
18
20
Distance/kb
0.0000
0.0088
0.0176
Map units

Since no appreciable pairwise disequilibrium was observed except between four closely-linked sites, three-way associations were not determined for the data from this region. The outline relationship of disequilibrium to distance is quantitatively similar to that observed by Miyashita and Langley (1988) for the *white* locus region in *D. melanogaster,* no data points showing any marked departure from the overall trend which would invite a selectionist interpretation; it may be concluded that levels of multilocus association in these regions behave exactly as predicted by standard population genetics theory and decay rapidly to zero for distances exceeding 1-2 kb.

4.5. Comparison of the two regions

From the raw data obtained here (see Appendix IV), there is a suggestion that the *rosy-Ace* region may not only have a higher heterozygosity than the *achaete-scute* complex but also, curiously, that the number of cleavage sites for the enzymes used may be more frequent. Even within the *rosy* sequence there is marked variation ranging from only 11 recognition sites (*Rsa*I) to 33 (*Alu*I). The base composition of the DNA may be expected to vary between loci as well as being slightly biased towards A+T overall, and thereby have a noticeable effect upon the number of sites screened by the enzymes. However, this should not have affected the estimates of variability. Two independent estimates now exist of heterozygosity at *rosy* (π=0.003, Aquadro *et al.* 1988; π=0.014 and θ=0.007 5, this study) and four of heterozygosity at *ac-sc* (π=0.003 3 and θ=0.002 4, Beech and Leigh Brown 1989; π=0.002 2 and θ=0.001 8, Eanes, Labate and Ajioka, 1989; π=0.000 6 and θ=0.001 3, Aguadé *et al.* 1989; π=0.002 4 and θ=0.001 4, this study). The wide discrepancy in some of the π values, coupled with the high variances attached to both estimators precludes any definitive statement on possible differences in nucleotide variability between the two regions.

The contrasting patterns of nonrandom association in the two regions are an emphatic demonstration of their differing recombinational properties: at *ac-sc,* no form of relationship of disequilibrium to physical separation can be discerned, while at *rosy* both r^2 and D' can be seen to obey an inverse function of distance and much of the variance of the regression of r^2 on distance can be reduced by postulating an exponential type of decline. This

has biological meaning in terms of the process primarily responsible for the decay of associations, namely the asymptotic randomizing force of crossing-over whose probability increases with distance.

CHAPTER 5

DISCUSSION

Following the establishment of the working theoretical basis of linkage disequilibrium (Geiringer, 1944; Bennett, 1954; Lewontin and Kojima, 1960) a foundation was laid for the quantitative treatment of evolution and natural selection at the multilocus level. In the same way that consistently intermediate frequencies of an allele at a single locus can be cited as evidence for some form of balancing selection (either frequency-dependent or heterotic) at that locus, so the presence of strong nonrandom associations of alleles at different loci consistently across populations may be interpreted as evidence for natural selection maintaining these allele combinations, since we expect disequilibrium to be eliminated by recombination.

Equally, though, the experimental proof of such an assertion is subject to the need to rule out the numerous other mechanisms which could be responsible for disequilibrium, namely drift, migration and hitchhiking. The latter is, in practice, inseparable from selection and must be considered as a component of it, while the random element of genetic drift makes it unlikely to create disequilibria of the same sign and magnitude in all populations (Lewontin, 1974).

The relative importance of these factors may depend to a large extent on the species studied: many organisms have a population structure or mating system inherently more conducive to the persistence of nonrandom associations, with or without selection. The habitual inbreeding and asexual propagation of many plant and protozoon species must especially be taken into account, while some mammals are characterized by rather small, subdivided populations in which the influence of drift and migration cannot be ignored. In spite of these reservations, many experimental studies of variation have been undertaken over the last 20 years in which attempts to detect linkage disequilibrium have been motivated largely by the desire to identify multilocus selection.

5.1. Early observations of disequilibrium

5.1.1. Morphological characters

The existence of functionally related polymorphic characters which combine in such a way as to produce discrete conserved phenotypes implies that certain phenotypic associations are maintained by selective pressures. The well-known phenomenon of heterostyly in the primrose, *Primula* involves, among others, two characters, style length and anther length, for which only two phenotypes are common in nature: long style/short anthers ('pin') and short style/long anthers, or 'thrum' (Mather, 1950; Dowrick, 1956). This strategy optimizes the chance of outbreeding, the former acting as an effective female and the latter an effective male flower. The mechanism is enhanced by a physiological self-incompatibility: crosses between two pin or two thrum flowers are lower in fertility than pin-thrum crosses, ensuring that the two types remain at approximately equal frequencies in the population. The entire set of characters segregates as though it were controlled by a single locus with thrum as the heterozygote and pin as the recessive homozygote; hence the genes in the complex must interact strongly with respect to the reproductive component of fitness.

Another well-documented example of such interaction concerns the patterns on the shells of snails such as *Cepaea nemoralis*. Here there is variation for the background colour between pink and yellow, and also for presence and absence of banding; although complex, these phenotypes are observed more often in certain combinations or 'morphs' which differ between habitats (Cain and Sheppard, 1954) and probably serve to achieve maximum crypsis on different substrates. As the fitness component in this case is due to predation pressure, there is the additional possibility of frequency-dependent selection for the rarer morphs. Attempts have been made to calculate disequilibrium values for these loci using, as a measure of phenotypic association, the statistic

$$Z_p = N_{12}N_{21}/N_{11}N_{22}$$

where N is the number in each phenotypic class, the subscripts 1 and 2 referring to the dominant and recessive alleles at each locus respectively

(Hedrick *et al.*, 1978). With random association of phenotypes Z_p=1, but for the colour and banding loci in *C. nemoralis* Z_p was found to average 2.3 (Clarke, B. *et al.*, 1968) and 1.6 (Carter, 1968) indicating a significant association of pink with bandedness, and yellow with lack of banding. A similar conclusion has been reached for the equivalent loci in *C. hortensis*, where *D'* was estimated as unity in 27/66 samples and positive in another 30 (Bantock and Noble, 1973).

The complex colours and morphology influencing mimicry in butterflies have also been shown to display interaction for fitness. Female *Papilio memnon* are known to mimic several other species, only some of which have conspicuous 'tail' lobes on the wings. Correspondingly there exist two mimetic forms of *P. memnon*, tailed and tailless, a polymorphism which is under the control of at least three loci with multiple alleles (Clarke, C.A., *et al.*, 1968). Occasional putative recombinants are observed with the wrong combination of colour and wing morphology; these are rare and presumed to be much lower in viability. Hence the population overall must display submaximal linkage disequilibrium of a highly adaptive nature.

Such gene complexes as these may be justifiably termed 'supergenes' (Darlington and Mather, 1949) since the genotype at each of the component loci is less important than the combined genotype in determining fitness. The inevitable selection for tighter linkage of such loci (Kimura, 1956) may result in opportunistic chromosomal rearrangements which reduce recombination between them becoming incorporated as permanent features of many supergenes.

5.1.2. Gene complexes: the *HLA* system

Another, quite distinct, category of 'supergene' is defined by the so-called multigene families such as the *HLA* (human leukocyte antigen), *Rh* and *ABO* blood group loci in man, or the *H-2* complex of the mouse. The tight linkage of the genes in these clusters stems from their common ancestry, rather than from selection as in the classical supergene: they are thought to have arisen by a series of tandem duplications followed by diversification. Invariably there is extensive polymorphism within such complexes which can be considered to have an adaptive function in terms of the genesis of diversity in the cell surface antigens, making them ideal

systems to examine for possible selectively-maintained linkage disequilibria.

The *HLA* system is a particularly well-studied multigene family consisting of over 1 500 loci in a tightly-linked cluster within 2.5 map units on chromosome 6, coding for the class I and II antigens of the human major histocompatibility complex. These cell surface molecules are targets for the T-lymphocytes that mediate self-recognition in the immune response; the system is one of the most polymorphic known in man, with typically 10-40 alleles per locus being expressed in the human population (see Auffray and Strominger, 1986).

Strong nonrandom associations were observed from the outset among these loci, both between serotypes and between serologically-determined markers and disease phenotypes (reviewed by Bodmer and Bodmer, 1978). Associations are especially common between alleles of the *HLA-B* and *-C* antigenic subgroups, where they are also consistent across several racial groups. Many of the disease associations involve various types of autoimmune disorder such as rheumatoid arthritis and juvenile-onset diabetes. These combined features of high heterozygosity with multiple alleles and pronounced disequilibrium would not normally be expected to occur in the absence of natural selection.

Neutral theory predicts a distribution of alleles which is heavily skewed, with relatively few alleles at intermediate frequencies (Ewens, 1972); this finding enables a test of fit to neutrality expectations to be made by comparing observed Hardy-Weinberg homozygosities with those expected under a neutral model (Watterson, 1978). Application of this approach to *HLA* serotypes from a diverse set of populations revealed significantly less homozygosity than could be accommodated by the neutral model for the class I and II loci (Hedrick and Thomson, 1983; Klitz *et al.,* 1986). Hence the inference of balancing selection from the observation of extensive single-locus polymorphism at *HLA* appears to be justified, and it is highly plausible that any such selection would involve a frequency-dependent component.

On the other hand, the frequent observations of linkage disequilibrium across the *HLA* complex need to be considered with more caution. The correlation of certain haplotypes with disease implies that some combinations may have been cemented either by selection or hitchhiking, but in this case

the common ancestry of the genes involved must be weighed against any such interpretation: the observed associations are probably due at least in part to the recent history of the duplications. The resultant sampling of entire multilocus genotypes by genetic drift would be exacerbated by the tight linkage of the daughter elements ensuring that these 'founder haplotypes' were not rapidly dissipated by recombination.

5.1.3. Inversions and allozyme loci

The first observations of linkage disequilibrium at the molecular level were between third-chromosome inversions in *Drosophila pseudoobscura* and *D. persimilis,* and particular allozyme electromorphs for three loci within the inverted segment (Prakash and Lewontin 1968, 1971). Prior to this discovery, gametic associations between syntenic inversions in numerous *Drosophila* and other dipteran species had been demonstrated (Levitan, 1958; Brncic, 1961). These correlated blocks of variants became known as 'coadapted gene complexes' (Dobzhansky, 1970), implicit in this phraseology being the co-selection of large arrays of alleles maintained in coupling disequilibrium by their association with the inversion.

One cannot, however, necessarily attribute either of these phenomena to selective interactions since such large-scale rearrangements, in common with the tandem duplications generating multigene families, have the effect of sampling a unique set of alleles which is subsequently conserved, to an extent, by the concomitant suppression of recombination. Hence the force responsible for these disequilibria may be again a special case of genetic drift. Moreover, natural selection acting on individual loci will automatically create disequilibrium between neighbouring loci *via* the hitchhiking effect (Thomson, 1977) which will be allowed to reach quite high levels in the absence of recombination. Since inversions tend to suppress recombination over the entire chromosome arm, the same argument can be advanced for the commonly-observed associations between inversions and allozyme polymorphisms close to but outwith the inversion (Kojima *et al.,* 1970; Mukai *et al.,* 1971; Prakash and Levitan, 1973; Langley *et al.,* 1974).

5.2. Models of multilocus evolution

The development of mathematical treatments considering more than one locus enabled a theoretical determination of the expected multilocus associations arising from different models of interactive selection. Two classes of fitness interaction may be considered: additive and multiplicative. The former is a deviation of the fitness of the multilocus genotype from the sum of the fitnesses of the component single-locus genotypes, the latter a deviation from their product. Both types of interaction can be envisaged in different situations in nature: two genes influencing a quantitative character such as bristle score would be expected to contribute additively to the phenotype if there were no interaction, whereas multiplicative effects on fitness might be a feature of, say, genes controlling consecutive stages in a developmental pathway.

Most early models dealt with heterotic, multiplicative selection (*e.g.* Lewontin and Kojima, 1960) and showed that stable linkage disequilibrium could be attained even from an initial random combination if interactive selection was strong enough. Extension of these models to five loci (Lewontin, 1964) and 36 loci (Franklin and Lewontin, 1970) gave the first indications that linkage disequilibrium could be a major feature of the genome with diverse and dispersed loci segregating in fixed combinations. The idea that local disequilibria could act as 'seeds' for the accumulation of a chain of associations ultimately tying up the entire chromosome as a unit of selection (Slatkin, 1972) was a compelling one and stimulated a flurry of experimental activity aimed at testing its validity.

5.3. Test of the theory: allozyme data

If the forecasts of these theoretical studies were accurate, we would expect linkage disequilibrium between polymorphic loci in populations to be widespread, and not just confined to tightly linked gene complexes such as *HLA*. Electrophoretic techniques provided the means to test this prediction, given the unprecedented levels of polymorphism detected (around one third of all loci studied) in the earliest surveys of allozymes in man (Harris, 1966) and *Drosophila pseudoobscura* (Lewontin and Hubby, 1966; see also Lewontin, 1974). These methods were now applied directly to the search for examples of nonrandom associations between (apparently) arbitrarily-chosen allozyme

loci in a range of animal and plant populations (see Barker, 1979, for a general review).

5.3.1. Plant populations

A substantial body of allozyme data exists for several plant populations (for review see Brown, 1979), and it is here that some of the most extensive linkage disequilibria have been discovered. However, these were almost exclusively in self-fertilizing species such as barley (Clegg *et al.,* 1972; Weir *et al.,* 1972) or oat (Allard *et al.,* 1972) where the lack of gene flow may be expected to retard the approach to linkage equilibrium in a manner analogous to linkage (Weir and Cockerham, 1973) and as such cannot be construed as evidence for stable, selectively-maintained disequilibria. The few studies conducted on outcrossing species such as the sea campion *Silene maritima* (Baker *et al.,* 1975) or lodgepole pine *Pinus contorta* (Hedrick and Thomson, 1986) failed to detect comparable levels of disequilibrium to those found in the inbreeders, implying that the mating system is a critical factor in the maintenance of these associations. A limited amount of disequilibrium has since been uncovered in *P. contorta* (Epperson and Allard, 1987) who found ten out of 123 pairwise comparisons to be significant in two populations, of which seven involved the tightly-linked loci *Got-1, Per-1* and *Per-II* for which c=0.007.

5.3.2. Microorganisms

Surveys of the enzymes produced by unicellular organisms have revealed some striking examples of linkage disequilibria. In a sample of 109 clones of *Escherichia coli,* only 98 distinct haplotypes were recorded for 20 enzymes, even though 18 of these were polymorphic (Selander and Levin, 1980) while in the protozoon parasite *Trypanosoma cruzi,* 43 haplotypes for 15 enzymes, 14 of which were polymorphic, were observed in a sample of 121 stocks again representing only a small proportion of the genotypes expected if alleles at these loci were associated randomly.

In these cases the predominantly asexual reproduction of the organisms provides effective linkage for the loci involved, as does the inbreeding habit of many higher plants, thereby retarding the decay of any nonrandom associations whether produced by drift or selection.

5.3.3. Outbreeding animals: molluscs and vertebrates

A few isolated instances of allozyme disequilibria have been recorded in outcrossing animal species (other than *Drosophila,* which will be discussed separately below). Mitton and Koehn (1973) found significant departure from the random expectation of allele combinations for the functionally similar aminopeptidase and leucine aminopeptidase loci in the blue mussel *Mytilus edulis,* which in the absence of obvious inbreeding or migration was attributed to fitness interaction between the two. Similar adaptive function was ascribed to the significant disequilibrium observed between alleles of the esterase-α and -β loci in five of 17 samples of the terrestrial salamander *Plethodon cinereus* (Webster, 1973), while in a rigorous study of 12 loci in the oviparous tooth carp *Fundulus heteroclitus,* only one pair out of the possible 66 comparisons showed significant genotypic correlations which were consistent across populations and year of study, suggesting a specific selection (Mitton and Koehn, 1975).

Although interesting as individual cases, these sporadic findings fall far short of demonstrating the amounts of multiple linkage disequilibrium required of the Franklin-Lewontin model. A genuine test of its predictions was provided by the data from large numbers of allozyme loci, arbitrarily-chosen but with known linkage relations, from fruit fly populations.

5.3.4. *Drosophila* allozyme loci

Among the many electrophoretic surveys of enzyme loci in *Drosophila* initiated since 1970, instances of significant disequilibria have been few and far between. Charlesworth and Charlesworth (1973) reported four out of 30 significant comparisons of third-chromosome loci covering 20 map units in *D. melanogaster,* while Zouros and Krimbas (1973) found one significant association between *Xdh* and *Ao* alleles in two populations of *D. subobscura.* Aside from these, most studies of natural populations recorded the same general outcome: linkage disequilibrium was common between allozymes and inversions (*e.g.* Loukos and Krimbas, 1975) but in comparisons involving pairs of allozyme loci, no more significant associations were usually found than would be expected by chance (Mukai *et al.,* 1971; Langley *et al.,* 1974; Langley *et al.,* 1978).

The incidence of disequilibrium was, however, somewhat higher for some laboratory-maintained populations (Birley, 1974; Langley *et al.*, 1978; Laurie-Ahlberg and Weir, 1979) presumably due to the compressed population sizes. Clegg *et al.* (1980) monitored linkage disequilibrium over time in laboratory populations of *D. melanogaster* and observed a rapid decay of all two-, three- and four-locus correlations to zero even faster than expected from the assumption of neutrality; these rates of decay were matched by theoretical simulations (Clegg, 1978) the outcomes of which were shown to be highly sensitive to the parameters chosen, generating stable disequilibrium only when $s^2/c>4$ where s is the coefficient of interactive selection. In the light of the observed experimental rates of decay, the ratio may not in reality approach this value for *Drosophila* populations.

It therefore seemed that no evidence for the congealment of chromosomes postulated by the Franklin-Lewontin model was to be found in the experimental data from natural populations. Much of the allozyme data for *D. melanogaster* has been summarized by Langley (1977) who concluded that little linkage disequilibrium was apparent but that since enzyme structure and hence mobility constituted phenotype, not genotype, a large fraction of the underlying variation might be overlooked. The problem of pooling of alleles also impairs the sensitivity of this approach as does the difficulty of obtaining enough polymorphic loci sufficiently closely linked; the advantages to be gained from a DNA-level survey were therefore overwhelming.

5.4. Molecular genetics

The ability to analyse the DNA strand directly was obviously of immense importance to the global study of genetic variation. The capacity for resolving sites within as well as between genes by the use of cloned probes gave the potential for the sensitive detection of variable sites in the sequence of nucleotides, which was necessary not only for more powerful tests of the predictions of neutral theory but also for the more rigorous elaboration of nonrandom associations. The initial surveys using restriction enzymes did indeed reveal greater levels of disequilibrium than had been generally found in allozyme studies, but the combined data from many such surveys, principally in *Drosophila* and man, has gradually brought forth a rather less consistent pattern of results.

5.4.1. Human gene complexes

It was in the β-globin gene complex of man that a population-genetic survey of variation using restriction enzymes was first attempted (Jeffreys, 1979) and this region has since been subject to intensive study at the DNA level. Here it is the spatial distribution of linkage disequilibrium which provokes most interest as it bears no simple relationship to the physical distance separating the polymorphisms. Two clusters of variants situated on either side of the β-globin gene were found to display strong disequilibrium within the two separate segments of 32 and 8 kb, yet haplotypes for these two clusters were randomly distributed with respect to one another (Antonarakis *et al.,* 1982; Kazazian *et al.,* 1984). As the intervening distance is only 9 kb, this has been taken as evidence for non-uniform recombination with a 'hotspot' for crossing-over being implicated in the vicinity of the δ-globin gene which lies between the two clusters (Chakravarti *et al.,* 1984).

The approach of Chakravarti *et al.* (1984) in using disequilibrium data to infer rates of recombination may be criticized on two grounds. First, their choice of estimator is inappropriate in this context, D' being preferable to r^2 as it is entirely independent of allele frequency (Hedrick, 1988). One cannot, for example, use a low value of r^2 as evidence for recombination if the corresponding D' value is at or near unity, as disequilibrium would still be maximal for the given allele frequencies. Secondly, their values for c were based on several questionable assumptions concerning population size and the expected value of r^2 (Weir and Hill, 1986). Nevertheless, similar observations of clumped distributions of disequilibria in other human genes such as *HLA* (Serjeantson *et al.,* 1986) and in the murine equivalent *H-2* complex (Steinmetz *et al.,* 1982) have led to the view being proposed that nonuniform recombination, with crossing-over largely confined to definable hotspots, is a particular feature of mammalian genomes (Bodmer, 1986).

Many of the DNA polymorphisms observed in human gene complexes are associated with specific diseases. In the β-globin gene cluster, for example, high correlations were found between certain β-thalassaemia mutations and characteristic restriction-site haplotypes (Kazazian *et al.,* 1984), while in the case of *HLA,* polymorphic restriction sites have been found which are again associated with autoimmune disease (Festenstein *et al.,* 1986).

Where linkage disequilibrium can be demonstrated between a polymorphic marker and a known disease locus, the polymorphism can be an invaluable aid in the detection of carriers of the defective gene; at the DNA level it is far more likely that variant sites will be discovered in close enough linkage for these purposes.

5.4.2. *Drosophila* genes

Initial reports of DNA-level variation in *Drosophila* began to uncover more frequent nonrandom associations than had been commonly found between allozyme polymorphisms: Langley *et al.* (1982) found significant associations between four high frequency variants across 9 kb in the *Adh* region of *D. melanogaster,* while Leigh Brown (1983) discovered two restriction-site and two small insertion polymorphisms in the 87A7 *heat-shock* locus which were in highly significant linkage disequilibrium over a distance of 20 kb. Further investigations of *Adh* extended the observations of disequilibrium at this locus (Cross and Birley, 1986; Aquadro *et al.,* 1986; Kreitman and Aguadé, 1986) although the last of these was confined to a shorter DNA segment of 3 kb.

In all of these cases the relationship of disequilibrium to distance was a tenuous one; all found examples of strong disequilibria between distant pairs of sites but not necessarily between sites in closer proximity. This may be attributable to the difference in age of some of the mutations: Kreitman (1983) has postulated a possible mutational hotspot within the *Adh* region from the spatial distribution of polymorphisms, which could indicate that some are of recent origin and still in residual disequilibrium with flanking sites.

Extension of the molecular genetic approach to other *Drosophila* samples has, however, contradicted the previous evidence and it now appears that such disequilibria were a property of the loci studied, rather than of the *Drosophila* genome as a whole: successive investigations of the *white, Notch,* and *rosy* loci of *D. melanogaster* failed to reveal any more significant associations than would be expected from chance (Langley and Aquadro, 1987; Schaeffer *et al.,* 1988; Aquadro *et al.,* 1988) even though in many cases variants in close proximity had been resolved. Moreover, even the high disequilibrium at *Adh* in *D. melanogaster* was not mirrored by the equivalent locus in *D. pseudoobscura* (Schaeffer *et al.,* 1987). A more detailed

investigation of the *white* locus (Miyashita and Langley, 1988) found most sites over 2 kb apart to be in approximate linkage equilibrium, an assertion upheld by the available data from *rosy* including those presented here.

5.5. Overview

With the progress of these surveys, several interesting features have emerged. The first is that estimates for polymorphism and heterozygosity have been consistently lower at the level of the DNA than at the level of the gene product; P is typically 0.4 in *D. melanogaster* and θ around 0.1 per locus when estimated from allozyme data (see Aquadro *et al.,* 1988) but more usually 0.015 and 0.006 respectively from restriction enzyme data (*e.g.* Langley *et al.,* 1982). This is not surprising in view of the high probability that, over the entire length of a polypeptide chain, even a low frequency of base substitutions would cause at least one amino-acid substitution. Provided that sufficient enzymes are used, however, restriction-site surveys provide the more fundamental estimates of heterozygosity.

Secondly, most insertion-deletion variation detected is at low frequency relative to the observable site polymorphism, consistent with the view that the former tends to be slightly deleterious. Further support for this hypothesis comes from the distribution of this sequence-length variation which suggests a lower frequency of large insertions on the X chromosome than on the autosomes, coupled with a pronounced clustering of these events in the neighbourhood of known transcripts where the chromatin structure may be unusually receptive to them (Beech and Leigh Brown, 1989).

The third feature of interest concerns the number of observations of linkage disequilibrium, which initially became more frequent with the advent of restriction enzyme mapping but then declined as a greater variety of loci were studied. The results presented here from the *ac-sc* and *ry* loci fall at opposite ends of this spectrum of observations. On the one hand, at *rosy* we have negligible disequilibrium except between sites separated by less than 1 kb, while at *ac-sc* there is highly significant disequilibrium over greater distances than have previously been examined at the DNA level, but approaching those between some allozymes (*e.g.* Epperson and Allard, 1987). Substantially higher disequilibrium exists over 80 kb at *ac-sc* than was observed over only 3 kb in a large number of comparisons of *Adh* variant sites for the same NC

population (Kreitman and Aguadé, 1986).

5.6. Features of the results

5.6.1. Disequilibrium values

In a region of the genome such as *ac-sc* where recombination is known to be reduced in frequency, one would expect to find higher than average levels of linkage disequilibrium. The data presented here and in Beech and Leigh Brown (1989) together with the recent work of Eanes, Labate and Ajioka (1989) constitute the first convincing demonstration of extensive linkage disequilibrium over such a large tract of DNA, between pairs of sites up to 86 kb apart in this case. Aguadé *et al.* (1989) also found significant disequilibrium in 3/10 comparisons extending over 80 kb of the region in the North Carolina population, although these were smaller in magnitude and involved the same three sites.

Similarly, the associations evident at the three-locus level are among the first statistically significant values to be found in an experimental sample. Fewer significant values at the three-locus level are to be expected since three-way disequilibrium is additional to the sum of the pairwise values and as a result either stronger interlocus interaction (Hastings, 1986) or higher sample sizes (Brown, 1975) are generally necessary for its demonstration. It is conceivable that still higher order disequilibria may be forthcoming from the data, but to uncover statistically significant four-way disequilibria may require impracticably large samples.

5.6.2. Heterozygosity estimates

Estimates of average heterozygosity were calculated here by two methods: that of Ewens *et al.* (1981) denoted by θ, and that of Nei and Tajima (1981) as π. Of these, θ has been the more widely quoted and Langley *et al.* (1988) suggested that estimates of θ for *Drosophila* loci were beginning to converge at around 0.008. The estimated θ=0.007 5 for the *rosy* region from these data compares with the numerous quoted figures for *Adh* (0.007, Aquadro *et al.* 1986; Cross and Birley, 1986); *Notch* (0.007, Schaeffer *et al.* 1988) and *Amy* (0.006, Langley *et al.* 1988) in *D. melanogaster,* but the estimate of π derived here (0.014) is considerably higher than a previous figure for *rosy*

(0.003, Aquadro *et al.* 1988). The estimates were confined to the *rosy* gene and as such may not be representative of the whole region, bearing in mind the reported two-fold higher heterozygosity across the *l*S12, *ry, snake* and *hsc*2 genes compared to the 5' flanking segment in *D. melanogaster* and even higher, 7-fold excess in the same direction in *D. simulans* (Aquadro *et al.*, 1988). On the other hand, the higher estimate may be simply due to the increased sensitivity of detection which accrues from the use of four-cutter enzymes. The spatial distribution of site polymorphisms in the *rosy* region is interesting; from the estimated sizes of the fragments and the known sequence information (see 4.4.1.1), it appears that only one out of six base changes occurs in a position which would give rise to an amino-acid substitution, in line with our expectations from neutral theory (Kimura, 1983).

Aguadé *et al.* (1989) claim to have found a dramatic reduction in nucleotide variability in the *ac-sc* region, obtaining θ=0.001 3 and π=0.000 6. While the value of θ derived in this study (0.001 4) shows good agreement with theirs and is consistent with some reduction of variation in the region, their remarkably low π value shows a very large discrepancy with that of 0.002 4 obtained here, and in the opposite direction relative to θ. Large differences between θ and π are normally only expected when there is a high proportion of rare variants in the sample, which may indicate some departure from neutrality or an unusually high mutation rate to selectively neutral alleles. However, the difference in π between the two surveys is more probably a reflection of the high variance associated with it when the allele frequency distribution is skewed. In addition, the difference in enzymes used may account for the higher estimate in this survey owing to the more sensitive detection of variation by four-cutter enzymes or any non-uniformity which might exist in the distribution of recognition sites. The only other report of such low heterozygosity in *D. melanogaster* was provided for the *G6pd* gene region (Eanes *et al.*, 1989) where θ=0.000 7 and π is again lower, at 0.000 35 for the North American population.

Although few definite conclusions can be drawn about precise levels of variability owing to the extremely high variances attached to the available estimators, an empirical variance is beginning to be established from the data so far accumulated, indicating a typical average heterozygosity of θ=0.006-0.008 for a *Drosophila* locus (*e.g. Notch, white, Adh* and *Amy*), with

current evidence suggesting a somewhat lower level of variability (θ=0.002–0.004) in the 87A7 *heat-shock* and *G6pd* genes and in the *yellow-achaete-scute* complex (see table in Eanes, Labate and Ajioka 1989).

5.7. Implications of the results

5.7.1. Distribution of disequilibrium

The failure to find significant amounts of linkage disequilibrium at many *Drosophila* loci, although surprising in the light of both theoretical predictions from a selectionist premiss and earlier experimental precedents, is nevertheless unremarkable from a neutralist standpoint. Most new mutations are likely to be either deleterious, neutral or very slightly selectively advantageous. The deleterious fraction will almost certainly never reach frequencies high enough for significant disequilibrium to be recordable; the remaining fraction will take so long to do so (thousands of generations even with heterotic selection coefficients of order 0.01; see Thomson, 1977) that in the intervening generations recombination will have broken up any nonrandom associations, even those involving quite closely-linked sites.

If this is an accurate scenario, we could regard the pattern of disequilibrium exhibited by the *white* and *rosy* loci as representing the basal level of multilocus association for a region of the genome with typical recombination fraction and minimal interactive selection. The much more extensive disequilibrium displayed by the *achaete-scute* complex could be accounted for by mutation in the absence of recombination between surviving polymorphisms, without necessarily invoking selection. This leaves the *Adh* locus in a curious category of its own in that such nonrandom associations as have been demonstrated are not related in simple fashion to recombination distance, suggesting either the recent origin and spread of a number of mutations or at least some nonrandomizing selection and/or pseudoselection (hitchhiking). The consistency of the *Adh*-F/S allozyme polymorphism across all populations surveyed creates a plausible rôle for hypotheses of selection and associated hitchhiking at this locus, several restriction-site polymorphisms being in linkage disequilibrium with the allozyme itself (Aquadro *et al.*, 1986).

5.7.2. Evolution of the *ac-sc* complex

The results presented from the *achaete-scute* survey, while easily explained by the lack of crossing-over at this locus, are nonetheless remarkable for showing an unprecedented level of linkage disequilibrium involving both pairs and triplets of sites across a stretch of DNA that includes several genes. The circumstances embodied in the 'crystallization' concept of Lewontin and others are thus more fulfilled by this region of the *Drosophila* genome than any other so far studied, with profound consequences for its evolution. Here, presumably, no evolutionary forces can operate on isolated polymorphisms without affecting their interactions with surrounding loci: the entire complex becomes the unit of selection.

Some insights into the evolutionary history of the region may be forthcoming from its haplotype diversity, which is strongly biased towards a few discrete combinations. It is notable that two of these (1 and 4) are exact complements of one another, and are both at high frequency in the sample. The existence of complementary haplotypes at high frequency is predicted by the theory of selectively balanced polymorphisms (Lewontin, 1974) and has also been observed experimentally at *Adh* (Cross and Birley, 1986) so it is tempting to infer the action of balancing selection across the *ac-sc* complex, maintaining a stable disequilibrium as envisaged by the Franklin-Lewontin model of multiplicative heterotic selection. Arguing against this, however, is the rather low overall heterozygosity in the region which would not be expected if heterosis were the rule. Nevertheless, the presence of 1 and 4 at high frequency with no intermediate classes is difficult to explain on a wholly neutralist basis. The remaining haplotype diversity is easily explained by the series of single mutational steps from 1 or 5 as shown in Fig. 25, without the need to incorporate recombination. Haplotypes 8 and 9, meanwhile, differ from 5 and 1 respectively only in the presence of insertion IV; this disjunct distribution of IV on three relatively distant haplotypes suggests a multiple origin for this polymorphism. Some interaction between the sites is suggested by the observation that only parts of the haplotypes appear to have evolved: most of the presumed recent mutations are confined to the proximal half of the region. The status of *ac-sc* as a type of 'supergene' makes it unlikely that its constituent elements are independent in their contribution to fitness, and selection may have had some rôle to play in preventing free

randomization of certain variant combinations.

5.8. Limits of the analysis

5.8.1. Autocorrelation of data

The set of pairwise disequilibrium values obtained in this survey, or in any other of its kind, cannot be considered independent of one another: if disequilibrium is observed between loci *A* and *B,* and between *B* and *C,* then any disequilibrium observed between *A* and *C* is to some extent predicted. It will be apparent from Tables 1 and 2 that of the 21 pairwise comparisons made between polymorphic sites at *ac-sc,* no more than six are independent in this sense, and only three can be made which do not share at least one site. For the *rosy* data the problem is even more acute, with 14 pairwise comparisons, up to five of which can be regarded as independent (Table 5). The effect of this can be seen in the graph of the regression (Figs 26 and 27) where the curve is undoubtedly made steeper by the autocorrelation of nine data points all involving the four most tightly-linked variants. Such autocorrelation as a function of linkage is certain to confound any attempt to quantify the relationship of disequilibrium to distance that does not incorporate a series of markers in both tightly-linked and progressively more widely-spaced pairs, as was available for the *ac-sc* region.

This problem is an unavoidable feature of the experimental approach which is severely limited by the number of polymorphisms which can be detected, and obliged to extract the maximum amount of information from the data available. While autocorrelation is not likely to have unduly distorted the general conclusions of this survey, it should always be taken into account in the design of experiments and the interpretation of limited data sets. For example, the three out of 10 significant disequilibria in the data of Aguadé *et al.* (1989) appear at first to be well in excess of chance expectations until it is noted that the disequilibria involve only different combinations of the same three sites. The need for independence of comparisons adds great value to the few comprehensive surveys on the scale undertaken by Miyashita and Langley (1988) in the *white* locus region.

5.8.2. Experimental technique

Although the more recent surveys of genetic variation have increasingly made use of restriction enzymes as their principal tool, only a few have adopted four-cutter enzymes as their major initiative (Kreitman and Aguadé, 1986; Miyashita and Langley, 1988; this study). The increased sensitivity of fragment detection afforded by the combined use of four-cutters and *poly*propenamide gel electrophoresis has not only allowed the more prolific discovery of variation but also encountered some of the inextricable difficulties in its recording. In many cases the fragments detected are close to or below the size limit for resolution by autoradiography, and this causes further obstacles when attempting to re-hybridize membranes consecutively (see 3.2.5). The use of enzymes has always been a compromise between resolution of fairly tightly linked sites and ease of extension to greater distances, so it is perhaps inevitable that the advantages to be gained from four-cutters in the former sense should be met by a trade-off in terms of reduced efficacy in the latter.

For certain enzymes, the plethora of recognition sites can lead to confusion when fragments of a similar size are generated by different mutational events, as witnessed by the 'family' of *Cfo*I variants described in this work. This is a potentially serious problem for the estimation of disequilibrium, as the superimposition of the unresolved classes would mask any true disequilibrium present. If fine-scale variation proves the rule, therefore, then this could be an alternative reason why discoveries of linkage disequilibrium at the DNA level have proved less common than expected in recent years.

The amount of fine-scale variation detected at the *white* locus using four-cutter enzymes, much of which could be identified as minor sequence-length changes (Miyashita and Langley, 1988) certainly allows for this possibility. However, it was equivalent surveys using six-cutter enzymes, including that of Langley and Aquadro (1987) which first established the lack of disequilibrium at such loci, and these would not be expected to give rise to multiple, superimposed variants. Moreover, the present report of polymorphism at *ac-sc* does in fact show much significant disequilibrium involving four-cutter sites, which would not have been expected if extensive pooling of variants had occurred. Hence although resolution may be limiting in a few

cases it cannot conceivably be responsible for the widespread absence of disequilibrium from experimental samples; it is safe to conclude that levels of parametric disequilibrium do indeed vary substantially between loci and between species.

The techniques common to the current surveys are likely to be a transient phase in the elucidation of DNA-level variation, and the time may soon be ripe for a shift of emphasis towards direct sequencing as the means to conduct this more rapidly and efficiently becomes available. In particular, the advent of the polymerase chain reaction is sure to make sequencing, at long last, practicable for large samples and will thence, by its ability to read the complete DNA message with all its associated variations, automatically take preference over the current methods which can only provide estimates of variability the accuracy of which are not yet known. Sequencing has always been the 'ideal' method for this purpose, limited only by practical considerations but may be finally expected to supersede the use of restriction enzymes in the near future.

5.9. Suggestions for further work

The combined results of this study and those of Beech and Leigh Brown (1989), Aguadé *et al.* (1989) and Eanes, Labate and Ajioka (1989) have produced a fairly comprehensive picture of the restriction map variation and nonrandom associations to be found in the *achaete-scute* complex. This is the first time that extensive disequilibrium has been demonstrated over such a large tract of DNA in *Drosophila,* reflecting the peculiar balance of evolutionary forces imposed upon the region by the suppression of recombination. It would now be of interest to investigate other loci adjacent to the telomeres or centromeres of chromosomes to find out whether the expected suppression of crossing-over in such regions can generate comparable levels of linkage disequilibrium to those observed here, and evidence for this has already been presented from the centromeric *vermilion* locus in *D. ananassae* (Stephan and Langley, 1989).

The results obtained from the *rosy* region have been sufficient to gain an insight into the contrasting pattern of disequilibrium to be found between it and the *ac-sc* complex, with disequilibrium here falling off with distance perhaps more markedly than would have been expected from theory but in

accordance with the outcomes of some other investigations (Aquadro *et al.*, 1988; Miyashita and Langley, 1988). Nevertheless, more work would be desirable in order to achieve the initial objective of the survey; in particular, screening of the *Ace* gene sequence for tightly linked variable sites would allow a comparison both between adjacent sites within the two gene loci and across the wide intervening distance.

The *ry-Ace* region is an especially suitable one on which to focus attention, with its substantial genetic, molecular and DNA sequence information. Parallel studies on other such loci, along the lines of those already initiated on the *white* and *Notch* regions would be a useful exercise to establish the statistical generality of the outcome. The *bithorax* complex is an obvious candidate: like *ac-sc,* it consists of a set of functionally related genes with a key rôle in development, arranged in close proximity on the chromosome. Its position downstream from *rosy* on chromosome III especially invites a comparison of the two to discover if the same genomic congestion observed in the *ac-sc* complex could be demonstrated in a similar supergene adjacent to the relative fluidity of the *rosy* gene. It would effectively be a control for position within a chromosome.

5.10. Current perspectives

The tendency for these surveys to concentrate on *Drosophila melanogaster* has perhaps detracted from the general applicability of the results obtained. On the other hand, this species is in many respects an ideal one for the problem (ideal as in gas) with its large population sizes and cosmopolitan distribution ensuring that most populations are probably highly outbred and in approximate Hardy-Weinberg equilibrium. One consequence of this is that factors other than natural selection cannot account so readily for observed instances of multilocus association in this species as, for example, in comparable mammalian surveys (Ohta, 1982).

While it is essential to investigate a range of species including plants, microorganisms and those animals less fashionable in genetics (see 5.3), each set of observations must be interpreted against a background of the particular population structure, mating system and unique selection pressures impinging upon the species, coupled with the practical considerations of the amount of genetic and molecular knowledge surrounding it. With this in mind, studies on

Drosophila melanogaster of the kind described here and emanating from the last seven years are likely to remain in the forefront of the experimental elucidation of genetic variation and linkage disequilibrium.

5.11. Conclusions

Linkage disequilibrium is an enigmatic indicator of the evolutionary history of populations. On the one hand, it is a uniquely useful guide for such purposes as detecting multilocus selection and hitchhiking, estimation of effective population sizes and population divergence, and inference of non-uniform recombination, to add to its obvious medical implications in the use of molecular markers for the tracking and diagnosis of human genetic disease. On the other, it is notoriously difficult to attribute exclusively to any single one of its possible causes, and this is reflected in the relatively few categorical conclusions which have been accrued from over two decades of its study. Inevitably each case must be treated on its merits, the number of possible confounding factors being taken into account in assessing the importance of observed disequilibria. This project was designed to minimize these variables in a number of ways. Choosing a large, single, relatively panmictic population avoided the effects of drift and migration, and extracting chromosomes immediately after collection of the flies ensured that they represented a truly natural population sample. Preparation of large samples allowed the sensitive detection of a large fraction of the disequilibria, and concentrating on loci previously cloned and analysed genetically enabled the regression on distance to be constructed.

Although a precise function for disequilibrium with recombination fraction could not be derived owing to the uncertainty in mapping estimates for the *ac-sc* region and the incomplete screening of the *ry-Ace* region, the results have nevertheless clearly demonstrated the difference in disequilibrium to be found at two loci with contrasting rates of recombination. Whether this difference is due entirely to differential recombination or whether selection and hitchhiking are also responsible to some degree will only become clear as more samples are studied. The empirical incidence of linkage disequilibrium is gradually becoming clear with the diversity of loci, populations and samples now under investigation, and the precedents set by the available data should make the interpretation of multilocus variation progressively easier.

ACKNOWLEDGEMENTS

I wish to thank **Dr A.J. Leigh Brown** for his patient and enthusiastic supervision of this project. My gratitude is also extended to the following people for their invaluable contributions:

Robin Beech, for advice on all aspects of molecular biology and for donation of some X chromosome lines;

Lila Rutherford, for making fly food;

Debbie Hall and **Nicky Fraser**, for able technical assistance;

Sir Frank, for the photographs;

Julia Davidson, **Petra zur Lage** and **Sarah Ross** for their encouragement, exchange of useful ideas and good humour;

and last, but not least, the many people who against all odds taught me how to use a computer.

1. Aguadé, M., *et al.* 1989. Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster.* Genetics **122,** 607.

2. Allard, R.W., *et al.* 1972. Evidence for coadaptation in *Avena barbata.* Proc. Natl. Acad. Sci. U.S.A. **69,** 3043.

3. Antonarakis, S.E., *et al.* 1982. Nonrandom association of polymorphic restriction sites in the β-globin gene cluster. Proc. Natl. Acad. Sci. U.S.A. **79,** 137.

4. Aquadro, C.F., *et al.* 1986. Molecular population genetics of the alcohol dehydrogenase gene region of *Drosophila melanogaster.* Genetics **114,** 1165.

5. Aquadro, C.F., *et al.* 1988. The *rosy* region of *Drosophila melanogaster* and *Drosophila simulans.* I. Contrasting levels of naturally occurring DNA restriction map variation and divergence. Genetics **119,** 875.

6. Auffray, C. and Strominger, J.L. 1986. Molecular genetics of the human major histocompatibility complex. In Advances in Human Genetics **15** (eds Harris & Hirschhorn), 197. Plenum Press.

7. Baker, J., *et al.* 1975. Genetic polymorphism in the bladder campion, *Silene maritima.* Biochem. Genet. **13,** 393.

8. Bantock, C.R. and Noble, K. 1973. Variation with altitude and habitat in *Cepaea hortensis* (Müll.). Zool. J. Linn. Soc. **53,** 237. Cited in Hedrick *et al.,* 1978.

9. Beech, R.N., 1987. Insertion-deletion variation in the DNA of three natural populations of *Drosophila melanogaster.* Ph.D Thesis, University of Edinburgh.

10. Beech, R.N. and Leigh Brown, A.J. 1989. Insertion-deletion variation at the *yellow-achaete-scute* region in two natural populations of *Drosophila melanogaster.* Genet. Res. **53,** 7.

11. Bender, W., *et al.* 1983. Chromosomal walking and jumping to isolate DNA from the *Ace* and *rosy* loci and the bithorax complex in *Drosophila melanogaster.* J. Mol. Biol. **168,** 17.

12. Bennett, J.H., 1954. On the theory of random mating. Ann. Eugen. **18,** 311.

13. Biessmann, H., 1985. Molecular analysis of the yellow gene (*y*) region of *Drosophila melanogaster.* Proc. Natl. Acad. Sci. U.S.A. **82,** 7369.

14. Bingham, P.M., *et al.* 1982. The molecular basis of P-M hybrid dysgenesis: the rôle of the P element, a P-strain-specific transposon family. Cell **29,** 995.

15. Birley, A.J., 1974. Multi-locus polymorphism and selection in a population of *Drosophila melanogaster.* I: Linkage disequilibrium on chromosome III. Heredity **32,** 122.

16. Bodmer, W.F., 1986. Human Genetics: The Molecular Challenge. Cold Spring Harbor Symp. Quant. Biol. **LI,** 1.

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

18. Boyer, H.W. and Roulland-Dussoix, D. 1969. A complementation analysis of the restriction and modification of DNA in *Escherichia coli.* J. Mol. Biol. **41,** 459.

19. Brncic, D., 1961. Non random association of inversions in *Drosophila pavani.* Genetics **46,** 401.

20. Brown, A.H.D., 1975. Sample sizes required to detect linkage disequilibrium between two or three loci. Theoret. Pop. Biol. **8,** 184.

21. - 1979. Enzyme polymorphism in plant populations. Theoret. Pop. Biol. **15,** 1.

22. Cain, A.J. and Sheppard, P.M. 1954. Natural selection in *Cepaea.* Genetics **39,** 89.

23. Campuzano, S., *et al.* 1985. Molecular genetics of the *achaete-scute* gene complex of *D. melanogaster.* Cell **40,** 327.

24. Carramolino, L., *et al.* 1982. DNA map of mutations at the *scute* locus of *Drosophila melanogaster.* EMBO J. **1,** 1185.

25. Carter, M.A., 1968. Studies in *Cepaea.* II: Area effects and visual selection in *Cepaea nemoralis* (L.) and *Cepaea hortensis.* Philos. Trans. R. Soc. Lond. B. **253,** 397.

26. Cavalli-Sforza, L.L. and Bodmer, W.F. 1971. The Genetics of Human Populations. Freeman.

27. Chakravarti, A., *et al.* 1984. Nonuniform recombination within the human β-globin gene cluster. Am. J. Hum. Genet. **36,** 1239.

28. Charlesworth, B. and Charlesworth, D. 1973. A study of linkage disequilibrium in populations of *Drosophila melanogaster.* Genetics **73,** 351.

29. Chia, W., *et al.* 1986. Molecular analysis of the *yellow* locus of *Drosophila*. EMBO J. **5**, 3597.

30. Chovnick, A., *et al.* 1976. Organisation of the *rosy* locus in *Drosophila melanogaster*: evidence for a control element adjacent to the xanthine dehydrogenase structural element. Genetics **84**, 233.

31. Church, G.M. and Gilbert, W. 1984. Genomic sequencing. Proc. Natl. Acad. Sci. U.S.A. **81**, 1991.

32. Clarke, B., *et al.* 1968. Studies on *Cepaea*. VI: The spatial and temporal distribution of phenotypes in a colony of *Cepaea nemoralis* (L.) Philos. Trans. R. Soc. London B **253**, 521.

33. Clarke, C.A., *et al.* 1968. The genetics of the mimetic butterfly *Papilio memnon* L. Philos. Trans. R. Soc. London B **254**, 37.

34. Clegg, M.T., 1978. Dynamics of correlated genetic systems. II: Simulation studies of chromosome segments under selection. Theoret. Pop. Biol. **13**, 1.

35. Clegg, M.T., *et al.* 1972. Is the gene the unit of selection? Evidence from two experimental plant populations. Proc. Natl. Acad. Sci. U.S.A. **69**, 2474.

36. Clegg, M.T., *et al.* 1980. Dynamics of correlated genetic systems. V: Rates of decay of linkage disequilibria in experimental populations of *Drosophila melanogaster*. Genetics **94**, 217.

37. Cross, S.R.H. and Birley, A.J. 1986. Restriction endonuclease variation in the *Adh* region in populations of *Drosophila melanogaster*. Biochem. Genet. **24**, 415.

38. Darlington, C.D. and Mather, K. 1949. The Elements of Genetics. Allen and Unwin.

39. Devereux, J., *et al.* 1984. A comprehensive set of sequence analysis programs for the VAX. Nucl. Acids Res. **12**, 387.

40. Dobzhansky, T., 1970. Genetics of the Evolutionary Process. Columbia.

41. Dowrick, V.P.J., 1956. Heterostyly and homostyly in *Primula obconica*. Heredity **10**, 219.

42. Dretzen, G., *et al.* 1981. A reliable method for the recovery of DNA fragments from agarose and acrylamide gels. Anal. Biochem. **112**, 295.

43. Dubinin, N.P., *et al.* 1937. Crossing-over between the genes *'yellow', 'achaete',* and *'scute'.* Drosophila Information Service **8,** 76.

44. Eanes, W.F., Labate, J. and Ajioka, J.W. 1989. Restriction-map variation with the *yellow-achaete-scute* region in five populations of *Drosophila melanogaster.* Mol. Biol. Evol. **6,** 492.

45. Eanes, W.F., *et al.* 1989. Restriction-map variation associated with the G6PD polymorphism in natural populations of *Drosophila melanogaster.* Mol. Biol. Evol. **6,** 384.

46. Engels, W.R., 1979. Extrachromosomal control of mutability in *Drosophila melanogaster.* Proc. Natl. Acad. Sci. U.S.A. **76,** 4011.

47. - 1981. Estimating genetic divergence and genetic variability with restriction endonucleases. Proc. Natl. Acad. Sci. U.S.A. **78,** 6329.

48. - 1985. A set of P cytotype balancer stocks. Drosophila Information Service **61,** 71.

49. Epperson, B.K. and Allard, R.W. 1987. Linkage disequilibrium between allozymes in natural populations of lodgepole pine. Genetics **115,** 341.

50. Ewens, W.J., 1972. The sampling theory of selectively neutral alleles. Theoret. Pop. Biol. **3,** 87.

51. Ewens, W.J., *et al.* 1981. Estimation of genetic variation at the DNA level from restriction endonuclease data. Proc. Natl. Acad. Sci. U.S.A. **78,** 3748.

52. Feinberg, A.P. and Vogelstein, B. 1983. A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. **132,** 6.

53. Festenstein, H., *et al.* 1986. New HLA DNA polymorphisms associated with autoimmune diseases. Nature **322,** 64.

54. Ford, E.B., 1971. Ecological Genetics. 3rd Edition. Chapman & Hall.

55. Franklin, I. and Lewontin, R.C. 1970. Is the gene the unit of selection? Genetics **65,** 707.

56. Frydenberg, O., 1963. Population studies of a lethal mutant in *Drosophila melanogaster.* I: Behaviour in populations with discrete generations. Hereditas **L,** 89.

57. García-Bellido, A., 1979. Genetic analysis of the *achaete-scute* system of *Drosophila melanogaster.* Genetics **91,** 491.

58. Gausz, J., *et al.* 1979. Genetic characterization of the 87C region of the third chromosome of *Drosophila melanogaster.* Genetics **93,** 917.

59. Gausz, J., *et al.* 1986. Molecular genetics of the *rosy-Ace* region of *Drosophila melanogaster.* Genetics **112,** 65.

60. Geiringer, H., 1944. On the probability theory of linkage in Mendelian heredity. Ann. of Math. Stat. **15,** 25.

61. Ghysen, A. and Dambly-Chaudière, C. 1989. Genesis of the *Drosophila* peripheral nervous system. Trends in Genet. **5,** 251.

62. Golding, G.B., 1984. The sampling distribution of linkage disequilibrium. Genetics **108,** 257.

63. Hall, L.M.C., *et al.* 1983. Transcripts, genes and bands in 315,000 base-pairs of *Drosophila* DNA. J. Mol. Biol. **169,** 83.

64. Hanahan, D., 1983. Studies on transformation of *Escherichia coli* with plasmids. J. Mol. Biol. **166,** 557.

65. Hardy, G.H., 1908. Mendelian proportions in a mixed population. Science **28,** 49.

66. Harris, H., 1966. Enzyme polymorphisms in man. Proc. R. Soc. Lond. B **164,** 298.

67. Hastings, A., 1986. Multilocus population genetics with weak epistasis. II. Equilibrium properties of multilocus models: what is the unit of selection? Genetics **112,** 157.

68. Hedrick, P.W., 1985. Genetics of Populations. Jones & Bartlett.

69. – 1987. Gametic disequilibrium measures: proceed with caution. Genetics **117,** 331.

70. – 1988. Inference of recombinational hotspots using gametic disequilibrium values. Heredity **60,** 435.

71. Hedrick, P.W. and Thomson, G. 1983. Evidence for balancing selection at HLA. Genetics **104,** 449.

72. – 1986. A two-locus neutrality test: application to humans, *E. coli* and lodgepole pine. Genetics **112,** 135.

73. Hedrick, P.W., *et al.* 1978. Multilocus systems in evolution. In Evolutionary Biology **11** (eds Hecht, Steere & Wallace), 101. Plenum Press.

74. Hill, W.G., 1974a. Disequilibrium among several linked neutral genes in finite population. I: Mean changes in disequilibrium. Theoret. Pop. Biol. **5,** 366.

75. - 1974b. Estimation of linkage disequilibrium in randomly mating populations. Heredity **33,** 229.

76. - 1981. Estimation of effective population size from data on linkage disequilibrium. Genet. Res. **38,** 209.

77. Hill, W.G. and Robertson, A. 1968. Linkage disequilibrium in finite populations. Theoret. Appl. Genet. **38,** 226.

78. Holmes, D.S. and Quigley, M. 1981. A rapid boiling method for the preparation of bacterial plasmids. Anal. Biochem. **114,** 193.

79. Hudson, R.R., 1982. Estimating genetic variability with restriction endonucleases. Genetics **100,** 711.

80. Jeffreys, A.J., 1979. DNA sequence variants in the $^{G}\gamma$-, $^{A}\gamma$-, δ- and β-globin genes of man. Cell **18,** 1.

81. Jennings, H.S., 1917. Numerical results of breeding with linkage. Genetics **2,** 97.

82. Jiménez, F. and Campos-Ortega, J.A. 1979. A region of the *Drosophila* genome necessary for CNS development. Nature **282,** 310.

83. - 1987. Genes in subdivision 1B of the *Drosophila melanogaster* X-chromosome and their influence on neural development. J. Neurogenet. **4,** 179.

84. Johnson, D.A., *et al.* 1984. Improved technique utilizing nonfat dry milk for analysis of proteins and nucleic acids transferred to nitrocellulose. Gene Anal. Techn. **1,** 3.

85. Karn, J., *et al.* 1980. Novel bacteriophage λ cloning vector. Proc. Natl. Acad. Sci. U.S.A. **77,** 5172.

86. Kazazian, H.H. Jr, *et al.* 1984. Quantification of the close association between DNA haplotypes and specific β-thalassaemia mutations in Mediterraneans. Nature **310,** 152.

87. Keith, T.P., *et al.* 1987. Sequence of the structural gene for xanthine dehydrogenase (*rosy* locus) in *Drosophila melanogaster.* Genetics **116,** 67.

88. Kidd, S., *et al.* 1983. The *Notch* locus of Drosophila melanogaster. Cell **34,** 421.

89. Kidwell, M.G., *et al.* 1977. Hybrid dysgenesis in *Drosophila melanogaster*: a syndrome of aberrant traits including mutation, sterility and male recombination. Genetics **86,** 813.

90. Kidwell, M.G., *et al.* 1981. Rapid unidirectional change of hybrid dysgenesis potential in *Drosophila.* J. Hered. **72,** 32.

91. Kimura, M., 1956. A model of a genetic system which leads to closer linkage by natural selection. Evolution **10,** 278.

92. – 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press.

93. Klitz, W., *et al.,* 1986. Contrasting evolutionary histories among tightly linked HLA loci. Am. J. Hum. Genet. **39,** 340.

94. Kojima, K., *et al.* 1970. A profile of *Drosophila* species' enzymes assayed by electrophoresis. I: Number of alleles, heterozygosities, and linkage disequilibrium in glucose-metabolizing systems and some other enzymes. Biochem. Genet. **4,** 627.

95. Kreitman, M., 1983. Nucleotide polymorphism in the alcohol dehydrogenase locus of *Drosophila melanogaster.* Nature **304,** 412.

96. Kreitman, M. and Aguadé, M. 1986. Genetic uniformity in two populations of *Drosophila melanogaster* as revealed by filter hybridization of four-nucleotide-recognizing restriction enzyme digests. Proc. Natl. Acad. Sci. U.S.A. **83,** 3562.

97. Lado, K.M., *et al.* 1987. Naturally-occurring restriction map variation in the *rosy* region of *Drosophila simulans*: contrasts with *D. melanogaster.* Unpublished preprint of Aquadro *et al.* (1988).

98. Langley, C.H., 1977. Nonrandom associations between allozymes in natural populations of *Drosophila melanogaster.* In 'Measuring Selection in Natural Populations' (eds Christiansen and Fenchel). Springer-Verlag.

99. Langley, C.H. and Aquadro, C.F. 1987. Restriction-map variation in natural populations of *Drosophila melanogaster: white*-locus region. Mol. Biol. Evol. **4,** 651.

100. Langley, C.H., *et al.* 1974. Linkage disequilibrium in natural populations of *Drosophila melanogaster.* Genetics **78,** 921.

101. Langley, C.H., *et al.* 1978. Analysis of linkage disequilibria between allozyme loci in natural populations of *Drosophila melanogaster.* Genet. Res. **32,** 215.

102. Langley, C.H., *et al.* 1982. Restriction map variation in the *Adh* region of *Drosophila.* Proc. Natl. Acad. Sci. U.S.A. **79,** 5631.

103. Langley, C.H., *et al.* 1988. Naturally occurring variation in the restriction map of the *Amy* region of *Drosophila melanogaster.* Genetics **119,** 619.

104. Laurie-Ahlberg, C.C. and Weir, B.S. 1979. Allozyme variation and linkage disequilibrium in some laboratory populations of *Drosophila melanogaster.* Genetics **92,** 1295.

105. Leigh Brown, A.J., 1983. Variation at the 87A heat shock locus in *Drosophila melanogaster.* Proc. Natl. Acad. Sci. U.S.A. **80,** 5350.

106. Levitan, M., 1958. Non random associations of inversions. Cold Spring Harbor Symp. Quant. Biol. **23,** 251.

107. Lewontin, R.C., 1964. The interaction of selection and linkage. I. General considerations: heterotic models. Genetics **49,** 49.

108. – 1974. The Genetic Basis of Evolutionary Change. Columbia.

109. Lewontin, R.C. and Kojima, K. 1960. The evolutionary dynamics of complex polymorphisms. Evolution **14,** 458.

110. Lewontin, R.C. and Hubby, J.L. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II: Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura.* Genetics **54,** 595.

111. Lindsley, D.L. and Grell, E.H. 1967. Genetic variations of *Drosophila melanogaster.* Carnegie Inst. of Washington.

112. Loukos, M. and Krimbas, C.B. 1975. The genetics of *Drosophila subobscura* populations. V: A study of linkage disequilibrium in natural populations between genes and inversions of the *E* chromosome. Genetics **80,** 331.

113. Maniatis, T., *et al.* 1978. The isolation of structural genes from libraries of eucaryotic DNA. Cell **15,** 687.

114. Maniatis, T., *et al.* 1982. Molecular Cloning: A Laboratory Manual. Cold Spring Harbor.

115. Mather, K., 1950. The genetic architecture of heterostyly in *Primula sinensis*. Evolution **4**, 340.

116. Maynard Smith, J. and Haigh, J. 1974. The hitch-hiking effect of a favourable gene. Genet. Res. **23**, 23.

117. McDonell, M.W., *et al.* 1977. Analysis of restriction fragments of T7 DNA and determination of molecular weights by electrophoresis in neutral and alkaline gels. J. Mol. Biol. **110**, 119.

118. Melton, D.A., *et al.* 1984. Efficient *in vitro* synthesis of biologically active RNA and RNA hybridization probes from plasmids containing a bacteriophage SP6 promoter. Nucl. Acids Res. **12**, 7035.

119. Mitton, J.B. and Koehn, R.K. 1973. Population genetics of marine pelecypods. III: Epistasis between functionally related isoenzymes of *Mytilus edulis*. Genetics **73**, 487.

120. – 1975. Genetic organization and adaptive response of allozymes to ecological variables in *Fundulus heteroclitus*. Genetics **79**, 97.

121. Miyashita, N. and Langley, C.H. 1988. Molecular and phenotypic variation of the *white* locus region in *Drosophila melanogaster*. Genetics **120**, 199.

122. Mukai, T., *et al.* 1971. Linkage disequilibrium in a local population of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U.S.A. **68**, 1065.

123. Nei, M. and Tajima, F. 1981. DNA polymorphism detectable by restriction endonucleases. Genetics **97**, 145.

124. Ohta, T., 1982. Linkage disequilibrium with the island model. Genetics **101**, 139.

125. Ohta, T. and Kimura, M. 1969. Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. Genetics **63**, 229.

126. Prakash, S. and Levitan, M. 1973. Associations of alleles of the esterase-1 locus with gene rearrangements of the left arm of the second chromosome in *Drosophila robusta*. Genetics **75**, 371.

127. Prakash, S. and Lewontin, R.C. 1968. A molecular approach to the study of genic heterozygosity in natural populations. III: Direct evidence of coadaptation in gene arrangements of *Drosophila*. Proc. Natl. Acad. Sci. U.S.A. **59**, 398.

128. - 1971. A molecular approach to the study of genic heterozygosity in natural populations. V: Further direct evidence of coadaptation in inversions of *Drosophila.* Genetics **69,** 405.

129. Rubin, G.M., *et al.* 1982. The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. Cell **29,** 987.

130. Rushlow, C.A., *et al.* 1984. Studies on the mechanism of heterochromatic position effect at the *rosy* locus of *Drosophila melanogaster.* Genetics **108,** 603.

131. Scalenghe, F., *et al.* 1981. Microdissection and cloning of DNA from a specific region of *Drosophila melanogaster* polytene chromosomes. Chromosoma **82,** 205.

132. Schaeffer, S.W., *et al.* 1987. Restriction-map variation in the alcohol dehydrogenase region of *Drosophila pseudoobscura.* Mol. Biol. Evol. **4,** 254.

133. Schaeffer, S.W., *et al.* 1988. Restriction-map variation in the *Notch* region of *Drosophila melanogaster.* Mol. Biol. Evol. **5,** 30.

134. Schalet, A., *et al.* 1964. Structural and phenotypic definition of the *rosy* cistron in *Drosophila melanogaster.* Genetics **50,** 1261.

135. Selander, R.K. and Levin, B.R. 1980. Genetic diversity and structure in *Escherichia coli* populations. Science **210,** 545.

136. Serjeantson, S.W., *et al.* 1986. HLA class II RFLPs are haplotype-specific. Cold Spring Harbor Symp. Quant. Biol. **LI,** 83.

137. Slatkin, M., 1972. On treating the chromosome as a unit of selection. Genetics **72,** 157.

138. Smit-McBride, Z., *et al.* 1988. Linkage disequilibrium in natural and experimental populations of *Drosophila melanogaster.* Genetics **120,** 1043.

139. Southern, E.M., 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98,** 503.

140. Spierer, P., *et al.* 1983. Molecular mapping of genetic and chromomeric units in *Drosophila melanogaster.* J. Mol. Biol. **168,** 35.

141. Steinmetz, M., *et al.* 1982. A molecular map of the immune response region from the major histocompatibility complex of the mouse. Nature **300,** 35.

142. Stephan, W. and Langley, C.H. 1989. Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila ananassae* populations. I. Contrasts between the *vermilion* and *forked* loci. Genetics **121,** 89.

143. Thirion, J.P. and Hofnung, M. 1972. On some genetic aspects of phage resistance in *E. coli* K12. Genetics **71,** 207.

144. Thomson, G., 1977. The effect of a selected locus on linked neutral loci. Genetics **85,** 753.

145. Thomson, G. and Baur, M.P. 1984. Third order linkage disequilibrium. Tissue Antigens **24,** 250.

146. Tibayrenc, M. and Ayala, F.J. 1988. Isozyme variability in *Trypanosoma cruzi,* the agent of Chagas' disease: genetical, taxonomical, and epidemiological significance. Evolution **42,** 277.

147. 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.

148. Wahl, G.M., *et al.* 1979. Efficient transfer of large DNA fragments from agarose gels to diazobenzyloxymethyl-paper and rapid hybridization using dextran sulfate. Proc. Natl. Acad. Sci. U.S.A. **76,** 3683.

149. Watterson, G.A., 1978. The homozygosity test of neutrality. Genetics **88,** 405.

150. Webster, T.P., 1973. Adaptive linkage disequilibrium between two esterase loci of a salamander. Proc. Natl. Acad. Sci. U.S.A. **70,** 1156.

151. Weinberg, W., 1908. Über den Nachweis der Vererbung beim Menschen. Jahresh. Verein f. vaterl. Naturk. in Württemburg **64,** 368. Cited in Hedrick, 1985.

152. Weir, B.S. and Cockerham, C.C. 1973. Mixed self and random mating at two loci. Genet. Res. **21,** 247.

153. - 1978. Testing hypotheses about linkage disequilibrium with multiple alleles. Genetics **88,** 633.

154. Weir, B.S. and Hill, W.G. 1986. Nonuniform recombination within the human β-globin gene cluster (Letter to the Editor). Am. J. Hum. Genet. **38,** 776.

155. Weir, B.S., *et al.* 1972. Analysis of complex allozyme polymorphisms in a barley population. Genetics **72,** 505.

156. Zouros, E. and Krimbas, C.B. 1973. Evidence for linkage disequilibrium maintained by selection in two natural populations of *Drosophila subobscura*. Genetics **73**, 659.

157. Zouros, E., *et al.* 1977. The effect of combining alleles into electrophoretic classes on detecting linkage disequilibrium. Genetics **85**, 543.

APPENDIX

I. Units and Nomenclature

SI units have been used throughout this thesis, with the kilogramme, metre and second as base units. Hence for units of volume the l, ml and µl have been replaced by the dm^3, cm^3 and mm^3 respectively. Radionuclide activities are quoted in MBq (1 Bq = 1 s^{-1} = 1/37 nCi). Solution concentrations are given either as molarities (M= mol dm^{-3}) or as percentages; all percentages are by mass unless otherwise stated. The terminology 'per cent w/v' was avoided since the concept of a percentage involving unlike quantities is a logical nonsense. Centrifugal forces are given as multiples of the standard gravity, *g* according to the following formula:

$$\text{Relative centrifugal force} = 4(3.1416)^2 rn^2 g/32.2$$

where r=radius in ft, n=revolutions per s. Peculiar units such as the kilobase (kb) are, of course, retained.

Similarly, chemical nomenclature has wherever possible followed the guidelines of the International Union of Pure and Applied Chemistry (IUPAC). Systematic names likely to be unfamiliar to the reader are accompanied by their trivial equivalents in parentheses; however, no attempt has been made to use systematic names for certain complex organic compounds such as sugars and amino acids.

References:

- Harrison, R.D. (ed.) 1982. Nuffield Advanced Science 'Book of Data'. Longman.
- Tennent, R.M. 1979. Science Data Book (7th Edition). Oliver and Boyd.
- Weast, R.C. (ed.) 1983. Handbook of Chemistry and Physics (64th Edition). CRC.
- Windholz, M. (ed.) 1976. The Merck Index (9th Edition). Merck & Co.

II. Solutions and media

L-broth

10 g dm^{-3} Tryptone (Oxoid)
5 g dm^{-3} yeast extract (Oxoid)
5 g dm^{-3} NaCl

L-agar

L-broth supplemented with 0.1% glucose and made 1% with agar.

L-top agar

L-broth made 10 mM Mg^{2+} and 0.5% with agar.

ψ-broth

20 g dm^{-3} Bacto Tryptone (Difco)
5 g dm^{-3} Bacto yeast extract (Difco)
10 mM Mg^{2+}

ψ-agar

ψ-broth containing 14 g dm^{-3} Bacto agar (Difco).

T-broth

10 g dm^{-3} Tryptone (Oxoid)
5 g dm^{-3} NaCl

Tris buffer

tris[Hydroxymethyl]methylammonium chloride (Tris-HCl) and *tris*[hydroxymethyl]methylamine (Tris-Base) mixed in appropriate proportions according to the desired pH.

1 M Tris pH 8.5: 0.7 M Tris-HCl
0.3 M Tris-Base

pH 8.0: 0.8 M Tris-HCl
0.2 M Tris-Base

pH 7.5: 0.9 M Tris-HCl
0.1 M Tris-Base

TE

10 mM Tris pH 8.0
1 mM EDTA pH 8.0

TMN

10 mM Tris pH 7.5
100 mM NaCl
10 mM Mg^{2+}

PSB

TMN made 0.05% with gelatine.

Trichloromethane/3-methylbutanol

96% by volume trichloromethane (chloroform)
4% by volume 3-methylbutanol (isoamyl alcohol)

Buffered phenol

Phenol redistilled from solid and stored at -20 °C in the dark before being melted at 80 °C, saturated with water and the solution made 0.1% quinolin-8-ol (8-hydroxyquinoline), adjusted to pH 8.0 with 2 M Tris-Base and stored at 4 °C.

Phenol-trichloromethane

A 50:50 mixture of buffered phenol and trichloromethane/3-methylbutanol each prepared as described above.

TBE

89 mM Tris-Base
89 mM boric acid H_3BO_3
3 mM EDTA

20 x SSC

3 M NaCl
0.3 M *tri*sodium 2-hydroxypropane-1,2,3-tricarboxylate (*tri*sodium citrate)

SSPE

150 mM NaCl
10 mM NaH_2PO_4
1 mM EDTA

Ligase buffer

50 mM Tris pH 7.5
10 mM $MgCl_2$
10 mM *threo*-1,4-dithiolbutane-2,3-diol (dithiothreitol, DTT)
1 mM ATP.

Tfbl

30 mM potassium ethanoate
100 mM RbCl
10 mM $CaCl_2$
50 mM $MnCl_2$
15% by volume propane-1,2,3-triol (glycerol)
Adjust to pH 5.8 with 0.2 M ethanoic acid and sterilize by filtration.

Tfbll

10 mM piperazine-1,4-*bis*[ethanesulphonic acid] (PIPES)
10 mM RbCl
75 mM $CaCl_2$
15% by volume propane-1,2,3-triol
Adjust pH to 6.5 with KOH and sterilize by filtration.

STET

8% sucrose
0.5% by volume 4-[1,1,3,3-tetramethylbutyl]phenoxy*poly*ethoxyethanol (Triton-X 100, Sigma)
50 mM EDTA pH 8.0
50 mM Tris pH 8.0

Solution III

3 M K^+
5 M $CH_3CO_2^-$

OLB

0.1 mM dATP
0.1 mM dGTP
0.1 mM dTTP
25 mM $MgCl_2$
250 mM Tris pH 8.0
50 mM 2-thiolethanol (2-mercaptoethanol)
1 M 4-[2-hydroxyethyl]piperazineethanesulphonic acid (HEPES) pH 6.6
27 O.D. units cm^{-3} mixed hexadeoxyribonucleotides (Pharmacia)

Transcription buffer

40 mM Tris-HCl
6 mM $MgCl_2$
5 mM *erythro*-1,4-dithiolbutane-2,3-diol (dithioerythritol, DTE)
4 mM N-[3-aminopropyl]butane-1,4-diamine (spermidine)
pH 7.2.

Sodium phosphate buffer

1 M sodium phosphate buffer pH 7.2: 0.72 M Na_2HPO_4
0.28 M NaH_2PO_4

III. Fly lines

The following table shows the numbering of fly lines in this thesis (left) aligned with the corresponding numbering system in published papers (Beech and Leigh Brown, 1989; Macpherson *et al.*, in preparation) on the right.

Extracted X lines	
NC 1	1
3	2
4	3
8	4
11	5
12	6
14	7
15	8
17	9
18	10
19	11
21	12
24	13
25	14
26	15
28	16
31	17
33	18
34	19
35	20
36	21
38	22
41	23
50	24
51	25
57	26
76	27
27	28
37	29
47	30
49	31
54	32
60	33
92	34
93	35
94	36
100	37
103	38
104	40
127	41
138	42
152	43
158	44

Extracted III lines

Line numbers used in this thesis are shown in the left-hand column; lines carrying recessive lethals or sublethals on chromosome III are denoted with an asterisk (*). The lines have also been numbered sequentially (right-hand column) for use in future documentation.

NC 1*	1
2	2
3	3
6	4
8	5
9	6
10	7
11*	8
12*	9
14	10
16	11
19*	12
22	13
23*	14
25*	15
34*	16
35*	17
36	18
38	19
41	20
46*	21
47*	22
49*	23
50	24
51	25
54*	26
56*	27
58*	28
61*	29
62	30
63*	31
64*	32
66	33
67*	34
69*	35
71*	36
72*	37
74	38
76*	39
77*	40
79	41
80*	42
88	43
89	44

Extracted III lines (contd)

NC 94*	45
95	46
99*	47
100	48
103*	49
104	50
114*	51
118	52
120	53
122	54
126	55
132	56
133*	57
134	58
135*	59
136*	60
137	61
139	62
140*	63
141	64
145	65
150*	66
151	67
152*	68
153	69
156*	70
157	71
158*	72

IV. Summary of data

Table A1. Fragment sizes and allele frequencies for the six restriction-site polymorphisms and one insertion event described in the *ac-sc* complex (section 3.1) as deduced by autoradiography.

Probe (Length/kb)	Enzyme (Polymorphism)	Fragment sizes/kb	Frequency
pASC101R5	MspI	1.7	0.16
(2.3)	(M55)	1.3 + 0.4	0.84
		0.85	1
		0.39	1
pASC133R1	CfoI	5	0.26
(1.9)	(C67)	2.1	0.74
pASC64R3	HaeIII	0.98	0.14
(1.9)	(H28)	0.75 + 0.23	0.86
		0.6	1
		0.3	1
pASC94R4	BglII	10	0.02
(3.3)	(G48)	7.2	0.98
		3.5	0.64
		2.1 + 1.4	0.36
sc31S2	BglII	11 (IV)	0.17
(7.6)	(G-19, IV)	6.8	0.83
		5.0	0.57
		3.3 + 1.7	0.43
sc22G2	XbaI	13.1	0.66
(5.5)	(X11)	11.7	0.34

Table A2. Fragment sizes and allele frequencies for the seven restriction-site polymorphisms described in the *rosy* gene region (section 3.2) as deduced by autoradiography.

Probe (Length/kb)	Enzyme (Polymorphism)	Fragment sizes/kb	Frequency
42-S3	BamHI	17.6	0.44
(1.9)	(B-178)	15.9 + 1.7	0.56
pRA-G3	CfoI	0.582	1
(3.4)	(C-167.3)	0.530	0.94
	(C-168.3)	0.507	0.92
	(C-169.3)	0.469	0.97
		0.449	1
	(C-167.3)	0.380 + 0.150	0.06
	(C-168.3)	0.345 + 0.162	0.08
	(C-169.3)	0.330 + 0.139	0.03
		0.326	1
		0.231	1
		0.158	1
	HaeIII	1.001	0.45
	(H-167)	0.566	1
		0.528 + 0.473	0.55
		0.447	1
		0.408	1
	AluI	0.833	1
	(A-167)	0.581	1
	(b)	0.359	0.08
	(a)	0.265	0.35
		0.250	0.56
		0.232	1

Table A3. Data used in the estimation of DNA sequence variability in the *ac-sc* region (see 4.3.1), as deduced by autoradiography, where m=no. of cleavage sites; k=no. of polymorphic cleavage sites; n=no. of lines surveyed.

N.B. Two of the polymorphisms included here, marked with hats, showed up very weakly and were not used in the disequilibrium calculations.

Opposite page 125

Probe	Enzyme	m	k	n	(q)
pASC133R1	CfoI	2	1	34	(0.26)
	MspI	4	1^	15	(0.2)
	MboI	2	0	16	
	TaqI	4	0	16	
	HaeIII	6	0	16	
	RsaI	4	0	10	
	AluI	5	0	8	
	BanII	3	0	32	
pASC101R7	BanII	2	0	30	
	MspI	3	0	12	
	MboI	3	0	14	
	RsaI	2	0	7	
pASC101R5	CfoI	7	0	8	
	MspI	5	1	36	(0.16)
	MboI	4	0	23	
	AluI	5	0	10	
	RsaI	4	0	10	
	HaeIII	7	0	5	
	BanII	2	0	30	
pASC53R1	AccII	3	0	16	
	MboI	7	0	14	
	MspI	6	0	12	
	CfoI	8	0	10	
	HaeIII	4	1^	10	(0.1)
	TaqI	9	0	5	
	RsaI	5	0	7	
	AluI	6	0	9	
	BanII	2	0	16	
pASC64R3	MboI	6	0	16	
	MspI	2	0	27	
	HaeIII	5	1	42	(0.14)
	TaqI	4	0	13	
	AluI	3	0	10	
	CfoI	3	0	10	
	RsaI	4	0	14	
	BanII	4	0	16	

$\Sigma m_4=143$ $\Sigma k_4=5$ $\Sigma m_5=13$ $\Sigma m=156$

Total Hardy-Weinberg homozygosity, $\Sigma(1 - 2pq)=3.6$.

Weighted value of $j = 4m_4/m + 5m_5/m = 4.08$.

Proportion of shared sites $S = (m-k+\Sigma[1-2pq])/m = 0.991$

Mean n=16.

Table A4. Data used in the estimation of DNA sequence variability in the *rosy* gene region (see 4.4.1.2), as deduced by autoradiography. Symbols as in Table A3 above. Three polymorphisms, marked with hats, are additional to those described in the text.

Probe	Enzyme	m	k	n (q)
pRA-G3/	AluI	5	2	62 (0.44,0.19)
4.6R/	CfoI	9	3	63 (0.03,0.08,0.06)
pRA-BR10	HaeIII	6	1	71 (0.45)
	RsaI	6	1^	32 (0.13)
	MboI	5	1^	32 (0.06)
	MspI	6	1^	22 (0.09)
	BanII	2	0	25

$\Sigma m_4=28$ $\Sigma k_4=9$ $\Sigma m_5=2$ $\Sigma m=39$

Total Hardy-Weinberg homozygosity $\Sigma(1 - 2pq) = 6.88$

Proportion of shared sites $S = (m-k+\Sigma[1-2pq])/m = 0.946$

Weighted value of $j = 4m_4/m + 5m_5/m = 4.05$

Mean n=44