MOLECULAR BIOLOGICAL STUDIES OF THE *TOPAZ* GENE REGION FROM THE SHEEP BLOWFLY, *LUCILIA CUPRINA*

A THESIS SUBMITTED TO THE
AUSTRALIAN NATIONAL UNIVERSITY
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
BY ABIGAIL ELIZUR-CHANDLER
FEBRUARY 1987

DECLARATION

The research described in this thesis was carried out by myself (except where due reference is made in the text), under the supervision of Dr. A.J. Howells.

Abigail Elizur-Chandler

ABIGAIL ELIZUR-CHANDLER

ACKNOWLEDGMENTS

I would like to express my gratitude and affection to my supervisor, Dr Tony Howells, for his guidance, enthusiasm and encouragement throughout the course of this research. Thanks also to my advisors Dr Liz Dennis and Dr Max Whitten for their advice and interest in my work, and to George Miklos, Ken Reed, Allan Lohe, John Oakshott and Rob Saint for stimulating discussions.

I would also like to express my appreciation to the blowfly group from CSIRO, Division of Entomology, in particular Geoff Clark, Geoff Foster and Dan Bedo for making their fly stocks freely available and for sharing their knowledge. The members of the Biochemistry Department have all been most helpful and friendly, making the years spent there very pleasant. Special thanks to all the people in Ken Reed's group who always generously shared their enzymes and vast technical experience. From the fly group, I would like to acknowledge the technical assistance of Ralli Klima, Vanessa Corrigan and Fiona Morris, and thank Donna Boyle, Janet Pagan and Cathy Landsberg for their close friendship and Barton Wicksteed for his sense of humour. My fellow Ph.D. students Mandy Walker, Patrick Skelly and Rick Tearle created an exciting working environment which shall be most fondly remembered.

To my dearest friend Peter, for his critical reading of this thesis and continual support while it was written, I am most indebted.

An Australian Wool Corporation award is gratefully acknowledged. Finally, many thanks to Pauline Kowacz for her accommodating attitude while expertly typing this thesis, and to Roland Aronsen for his assistance in illustration.

ABSTRACT

The *topaz* gene from the sheep blowfly, *Lucilia cuprina* was isolated and characterized by restriction enzyme and Southern blotting analysis as well as by DNA sequencing. In addition to the wild-type gene, two spontaneous *topaz* mutant alleles have been isolated and studied.

The wild-type *topaz* gene was isolated by exploiting the fact that *scarlet*, the homologous gene from *Drosophila melanogaster* has been recently cloned (Tearle, 1986). An initial experiment showed that *scarlet* sequences could probably be used to identify the *topaz* gene. DNA from a λ-clone carrying the *scarlet* gene region was used to probe a Southern blot of *L.cuprina* genomic DNA. A clear hybridization signal was obtained, indicating there is sequence homology between the *scarlet* gene region and sequences in *L.cuprina*, probably corresponding to the *topaz* gene.

A *L.cuprina* genomic DNA library in the λ-derived vector EMBL3A was prepared and screened with a fragment carrying sequences from the *scarlet* gene region. Seven recombinant phages were isolated; their restriction maps were determined and regions of homology to the *scarlet* gene region established. The restriction maps of six of the putative *topaz* clones fully overlap, while the seventh overlaps only partially. In addition to carrying sequences homologous to *scarlet*, the phages contain repeated sequences which are dispersed in the *L.cuprina* genome and interspersed with unique sequences in the clone. DNA fragments from the λ-clones were subcloned into a plasmid vector. The largest subclone containing only unique sequences homologous to *scarlet* was 2.7 kb. This subclone was used in Southern blotting experiments to determine the genomic organization of the putative *topaz* gene, and revealed that there is restriction fragment polymorphism around the *topaz* gene region in the wild type *L.cuprina* population.

The regions of *topaz* sharing homology with *scarlet* have been sequenced and compared with the *scarlet* nucleotide sequence. This comparison enabled the prediction of the *topaz* (and *scarlet*) gene structure. Regions of open reading frame that showed a high degree of homology at the level of the deduced amino acid sequence, and which were flanked by exon/intron consensus junction sequences (Mount, 1982) were predicted to be exons. The intervening sequences, which usually contained termination codons in all three reading frames and showed no homology at the nucleotide or predicted amino acid level were assumed to be introns. This putative gene structure reveals that *topaz* has seven exons (although the position of exon 1 has been determined by hybridization only and has yet to be confirmed by DNA sequencing). The exons characterized so far in the two genes (*topaz* and *scarlet*) are identical in size and are over 80% homologous in their predicted amino acid sequence. The introns vary considerably between the two genes; they share no obvious sequence homology and the introns in *topaz* are generally significantly longer than those in *scarlet*. In addition, at least four different repeated sequences are located in the *topaz* introns. One of these (located in intron 5) consists of tandemly repeating copies of the simple sequence TCT$\frac{A}{G}$; the other repeats, (located in introns 1,5 and 6) consist of repeating copies of a more complex nucleotide sequence. The two genes are also different in their GC composition; *topaz* has a lower G+C content than *scarlet* (in both exons and introns) and a different codon usage, exhibiting a strong preference for codons ending with A or T compared with *scarlet* which has a strong preference for codons ending with G or C. Hydropathy plots derived from the deduced amino acid sequence of *topaz* indicate that the *topaz* protein has several strongly hydrophobic domains (including one typical membrane spanning region near the C-terminus) and also hydrophilic regions, one of which shows homology to the putative ATP binding site of a number of bacterial transport proteins (Ames, 1986).

Genomic DNA libraries were prepared from the two mutants, *topaz*^{1} and *topaz*^{2}, and screened with a subclone from the *topaz*$^+$ gene that contains unique sequences only. Four *topaz*^{1} and one *topaz*^{2} clones were recovered, and restriction enzyme mapping has shown that whereas the *topaz*^{1} clones were all overlapping and cover the complete gene, the *topaz*^{2} clone is missing sequences from the 3' end (exons 6 and 7). Three additional *topaz*^{2} genomic libraries were constructed and screened with a unique sequence probe containing exon 7, however no positively hybridizing phage were recovered, indicating that the region contains sequences which clone poorly. Further clones carrying the 5'-end of the *topaz*^{2} gene were obtained from these libraries but have not been characterized.

The clones of the mutant genes were studied by hybridization and partial nucleotide sequence analysis. When compared with the *topaz*$^+$ gene, a number of features are apparent: (1) There are single base substitutions and small deletions and insertions present between the wild-type and the mutant alleles, many of which occur in intron regions and are probably a reflection of natural sequence polymorphism at the locus. Three of these single base substitutions resulted in the creation of restriction sites in *topaz*^{1} that are not present in *topaz*$^+$. (11) There are two DNA insertions in the first intron of *topaz*^{1} relative to *topaz*$^+$. These insertions are located 22 bp apart and are composed of sequences which are repeated elsewhere in the *L.cuprina* genome. Additional genomic clones, containing regions homologous to these repeated sequences were isolated and partially characterized revealing that at least one of the insertions is in fact a composite element, i.e. it is composed of a number of repeated sequence units that are repeated to different extents and probably in different combinations elsewhere in the genome. *Topaz*^{2} also has an insertion in intron 1 which is partially homologous to the composite repeat found in *topaz*^{1}, however smaller in size. (111) In the exon regions, *topaz*^{1} has an 18 bp deletion which removes 6 amino acids without affecting the reading frame and one single base substitution that results in a conservative amino acid substitution. The 6 amino acid deletion seems to be the most likely explanation for the mutant phenotype. In *topaz*^{2}, where less sequence data has been obtained, one non-conservative amino acid substitution was observed.

These results have provided insights into the general organization of sequences in the *L.cuprina* genome and of the complex structure of some of its dispersed repeats. They have also shown that restriction fragment polymorphism is common in natural populations of *L.cuprina* and this is brought about both by variation in restriction sites (due to single base changes) and also to distance between sites (due, in large part, to changes in the position or length of repeated sequence elements). The putative amino acid sequence of the *topaz* protein has provided further insight about its function and supports biochemical observations which predicted that the *topaz* protein is membrane associated and involved in transport of pigments and pigment precursors (Summers *et al*., 1982). The analysis of the *topaz* mutants revealed the possible nature of the mutations; significantly, neither seems to have been caused by the insertion of a "*D.melanogaster* like" transposable element.

FREQUENTLY USED ABBREVIATIONS

bp	:	base pairs
EtBr	:	Ethidium Bromide
IAC	:	Isoamylalcohol/chloroform 1:24
kb	:	kilobase (pairs)
SWT	:	standard wild type line of *Lucilia cuprina*

TABLE OF CONTENTS

CHAPTER 3 : THE ISOLATION AND CHARACTERIZATION OF THE WILD TYPE TOPAZ GENE

CHAPTER 4 : THE ANALYSIS OF THE TOPAZ GENE REGION BY DNA SEQUENCING

CHAPTER 1

INTRODUCTION

The work described in this thesis is the first detailed molecular analysis of DNA sequences from the sheep blowfly, Lucilia cuprina. The genomic region studied contains the coding region for the eye colour gene, topaz, and the questions addressed concern the gene structure and function, and sources of spontaneous mutations in topaz region compared with the equivalent gene region in Drosophila melanogaster, scarlet as well as the general sequence organization of the genome in the blowfly. In the longer term it is hoped that the availability of a cloned and well-characterized eye colour gene, which could act as an ideal phenotypic marker will form the basis for development of a gene transformation system in L.cuprina, as part of an assessment of the potential impact of recombinant DNA technology in blowfly control.

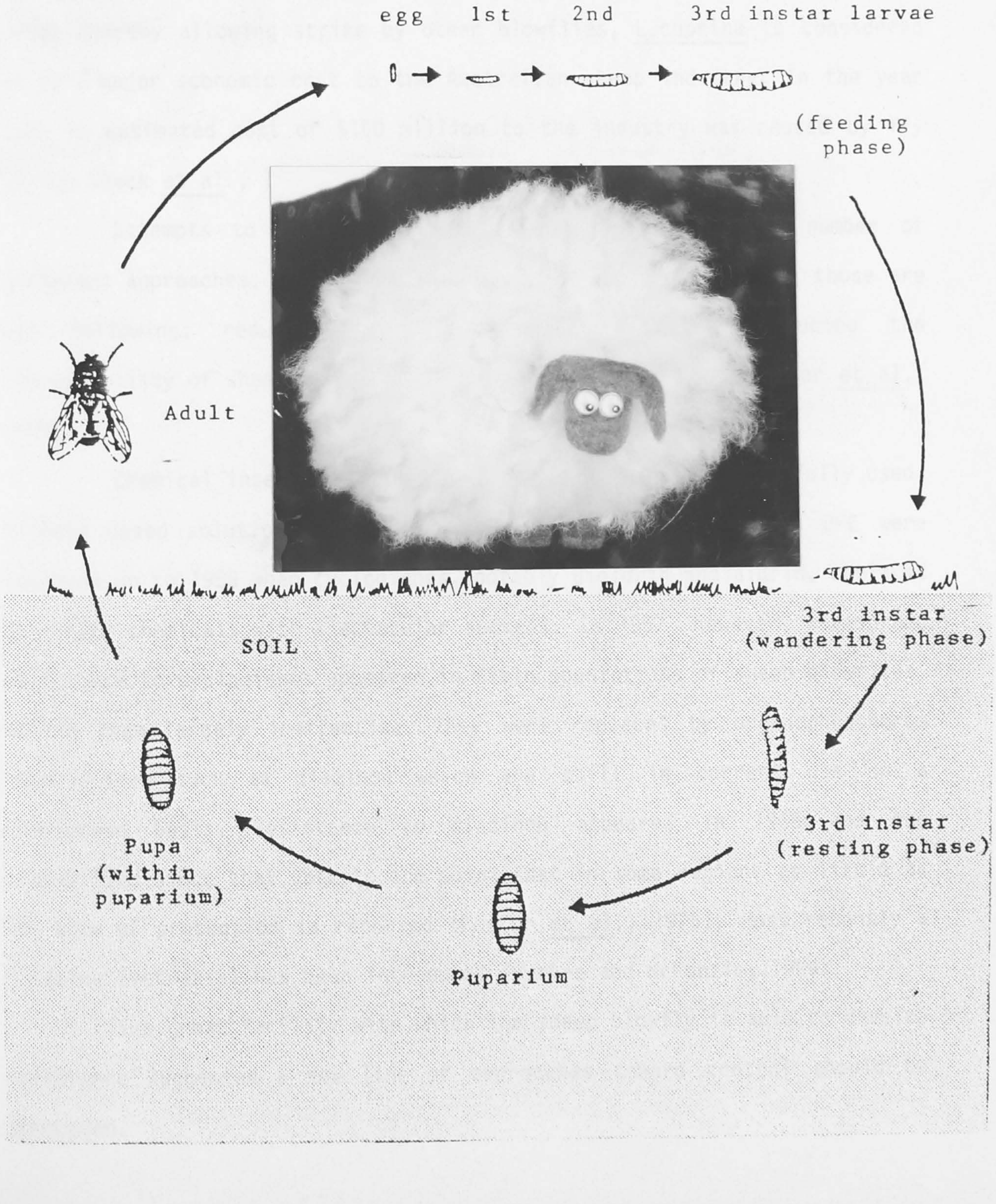
In this chapter the biology, genetics and eye pigmentation system of L.cuprina will be introduced, followed by a presentation of the general features of DNA sequence organization observed in the eukaryotic genome.

1.1 The biology of the sheep blowfly

The sheep blowfly, L.cuprina (Order Diptera, family Calliphoridae) is thought to have been introduced into Australia from South Africa or India in the 19th century. Originally the fly was a carrion feeder, however competition from native blowfly species for breeding space on carrion coupled with the introduction and farming of a large sheep population in Australia probably combined to promote its myiasis-causing behaviour (Foster and Whitten, 1974).

The life cycle of *L.cuprina* is shown in Figure 1.1. Female adults oviposite on areas of the sheep's body where there is superficial bacterial infection due to dampening of the skin by rain, sweat or urine. *L.cuprina* eggs hatch over a wide temperature range (10-40^{o}C) as long as they do not dry out. Young larvae feed on fluid exuding from the skin as a result of the superficial bacterial infection which is aggravated by their rasping mouthparts. Older larvae penetrate through the skin and attack living tissue causing extensive lesion around which secondary myiasis blowflies (mainly species of *Calliphora* and *Chrysomya*) lay eggs. The wound may be spread by repeated layings, which not only destroys or decreases the quality of meat and wool but can also give rise to a toxaemia which results in sheep death (Summers, 1979; Skelly, 1986). During the second and the early (feeding) phase of the third instar, larvae continue to feed and grow in the epidermis and dermis of the sheep. Late third instar larvae (wandering phase) drop off their hosts and eventually burrow into the soil where they assume a resting phase. Subsequently, a protective puparium forms (pupariation) within which metamorphosis and adult development ensues with the eventual emergence of the adult fly. Adults emerge during the early morning, they need water and carbohydrate as an energy source for sustained flight. In addition adult females require protein to become sexually receptive and to successfully mature their eggs. Depending on the ambient temperature, females in the field can mature a batch of eggs (potentially as many as 300) within 4-8 days. Females can survive several weeks in the field and at least five or six generations are possible in a season which can begin between August and November (depending on weather conditions) and end between April and June (Whitten *et al*., 1975; Skelly 1986).

FIGURE 1.1 The life cycle of the sheep blowfly, Lucilia cuprina (modified from Skelly, 1985).



1.2 Economic significance and control measures of *L.cuprina*

L.cuprina has been recognized as a major myiasis blowfly since early this century and is now known to cause at least 80% of primary strikes (Whitten *et al*., 1976). By initiating wounds in previously healthy sheep thereby allowing strike by other blowflies, *L.cuprina* is considered to be a major economic cost to the Australian sheep industry. In the year 1985 an estimated cost of $150 million to the industry was caused by fly strike (Beck *et al*., 1985).

Attempts to control blowfly strike have followed a number of different approaches, none of which are highly successful. Among those are the following; reducing fly densities by trapping, reducing the susceptibility of sheep and biological control (reviewed by Foster *et al*., 1975).

Chemical insecticides have been the weapon most successfully used. Arsenic based solutions and organic chemicals such as DDT and BHC were employed up to 1955 when cyclodienes, notably dieldrin and aldrin, replaced the other chemicals with spectacular success. In 1957, however, widespread resistance to cyclodienes appeared in field populations of sheep blowflies, making them largely ineffective. They were replaced by organophosphorus insecticides such as diazinon which are still in current use on a widespread basis. Resistance to diazinon appeared in 1965 and has intensified since that date to the point that in some regions as little as one week of protection is reported (Foster *et al*., 1975). Most recently a triazine (vetrazin) has been introduced (Hughes and McKenzie, 1986)

The great facility with which the sheep blowfly is able to develop resistance suggested a new line of approaches toward control should be attempted.

1.3 Genetic control strategies

The use of the insect itself in biological control (autocidal control) involves the release of insects which are sterile or genetically modified. When such flies are released in numbers much greater than the field population, the majority of matings will be with the released flies and will therefore be infertile or impose a high genetic load on subsequent generations. The sterile insect release method (SIRM) was used with initial success to control the screw-worm fly in southern United States and the mosquito in Florida (Bushland, 1971; Whitten and Foster, 1975).

The CSIRO Division of Entomology in Canberra has been studying the feasibility of using compound chromosome or Y;autosome translocation strains in the genetic control of L.cuprina. Due to the complicated genetics required for the development of compound chromosome strains, research has concentrated on the second alternative, the use of Y;autosome translocation strains carrying conditional lethal mutations on the translocated autosomes. These have low fertility when crossed with field populations (due to the translocated chromosomes) and their progeny have low viability due to the conditional lethal mutations. The conditional lethal mutations chosen for pilot studies result in eye colour abnormalities, which affect the fly's vision. White, topaz and yellow are all recessive mutations that were originally isolated for routine genetic analysis. When released in the field, flies with these mutations cannot be retrapped, indicating their lethality under field conditions although there is adequate viability and fertility under laboratory conditions (Whitten et al., 1976).

There are two major limitations to this system. The first concerns the male sterility which is associated with the translocation rearrangement, as partially sterile males create limitations in mass rearing the strain. The second limitation is that only conditional lethals which are located on the translocated chromosomes can be used.

1.4 The genetics of L.cuprina

Extensive genetic information is available regarding L.cuprina thanks to the research effort devoted by the CSIRO Division of Entomology to the study of the formal genetics and cytogenetics of the fly.

L.cuprina has a diploid complement of twelve chromosomes. These include five pairs of autosomes (numbered 2-6) and a single XX/XY, sex chromosome pair (Whitten et al., 1975). C-banding studies have shown the sex chromosomes to be heterochromatic and the autosomes each to have a small procentric C-banding region (Bedo, 1980). Giant polytene chromosomes corresponding to the autosomes can be prepared from the pupal bristle-forming trichogen cells (Foster et al., 1980). There are 130 known mutations of L.cuprina which have been genetically assigned to linkage groups (Maddern et al., 1986). Many of these mutations have been cytologically mapped to specific banded regions of the polytene chromosomes through the analysis of chromosome rearrangements (Foster et al., 1980; Foster et al., 1981).

Analysis of mutations in L.cuprina which have a biochemical or morphological phenotype similar to mutations observed in a number of other species of Diptera have shown the genetic linkage groups to be highly conserved (Foster et al., 1980; Foster et al., 1981). The apparent relationship between L.cuprina and D.melanogaster chromosome arms is shown in Table 1.1.

TABLE 1.1: POSSIBLE CORRELATIONS BETWEEN THE CHROMOSOMES OF L.CUPRINA AND D.MELANOGASTER (FOSTER ET AL., 1981)

L.cuprina	D.melanogaster
chromosome 2	chromosome 2L
chromosome 3	X-chromosome
chromosome 4	chromosome 3R
chromosome 5	chromosome 3L
chromosome 6	chromosome 2R

The availability of genetically characterized mutants affecting eye pigmentation in L.cuprina provided the basis for an extensive study elucidating the biochemical pathways involved in eye pigmentation in the blowfly (Summers 1979). This study supported the correlation between linkage groups in L.cuprina and D.melanogaster, and showed that the similarity between the two species extends to the biochemical level, at least with respect to eye pigment synthesis.

Access to cloned genes and to a transformation system could be a natural extension of the genetic work directed towards blowfly control because it would provide the opportunity of placing additional lethals on the translocated chromosomes as well as opening novel approaches to the problem (see review by Cockburn et al., 1984). However, before such work can be attempted, basic knowledge about the molecular biology of L.cuprina is required.

Since the study reported in this thesis concerns the isolation and characterization of the L.cuprina eye colour gene, topaz, the biology and biochemistry of the eye pigmentation system will be reviewed next.

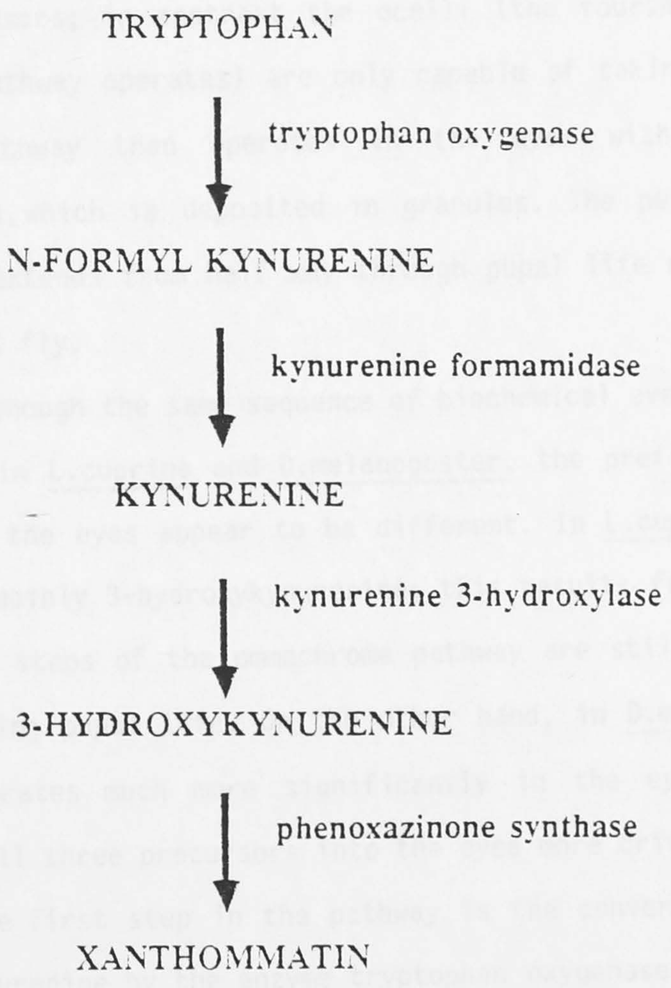
1.5 Eye pigmentation in L.cuprina

The wild type eye colour of the L.cuprina adult fly is orange-brown. This colour is due to the presence of two types of screening pigments in the pigment cells that act to optically isolate each facet (ommatidium) of the compound eye from its neighbours (Summers, 1979). There are two independent biochemical pathways that lead to the production of the major screening pigments, xanthommatin and sepiapterin. Xanthommatin, the brown pigment, is derived from tryptophan while sepiapterin is a yellow pteridine which derives from guanine (Summers et al., 1982). Since the work described in this thesis does not concern the pteridine pathway, it will not be further described.

1.6 The ommochrome biosynthetic pathway

The biochemical steps involved in xanthommatin production are shown in Figure 1.2. The pathway operates in four organs and in two phases during development in L.cuprina. During larval life, free tryptophan is taken up by the malpighian tubules and the larval fat body and stored as xanthommatin precursors. The pathway is only partially operative in these two organs, and the tryptophan is converted to kynurenine and 3-hydroxykynurenine in the malpighian tubules, or to kynurenine only in the fat body. Tryptophan itself is stored in the larvae in the form of larval proteins. In both L.cuprina and D.melanogaster about 30% of the tryptophan content of the fully-grown larvae finishes up as xanthommatin in the adult eyes.

FIGURE 1.2 The xanthommatin biosynthetic pathway in insects (from Summers, 1979).



Soon after pupariation the larval fat body begins to degrade releasing into the haemolymph tryptophan (from the proteolysis of stored proteins) and kynurenine. The malpighian tubules also release tryptophan, kynurenine and 3-hydroxykynurenine at this time. At about half way through pupal life, the pigment cells in the eyes gain the capacity to take up all three precursors; in contrast the ocelli (the fourth group of organs in which the pathway operates) are only capable of taking up kynurenine. The complete pathway then operates in the eyes with the production of xanthommatin, which is deposited in granules. The period of xanthommatin deposition extends from half way through pupal life until after emergence of the adult fly.

Although the same sequence of biochemical events (described above) occur both in L.cuprina and D.melanogaster, the preferences for precursor uptake into the eyes appear to be different. In L.cuprina the eyes appear to take up mainly 3-hydroxykynurenine; this results from the fact that the first three steps of the ommochrome pathway are still active in the body tissues during pupal life. On the other hand, in D.melanogaster, the full pathway operates much more significantly in the eyes, making efficient uptake of all three precursors into the eyes more critical (Summers, 1979).

The first step in the pathway is the conversion of tryptophan to N-formylkynurenine by the enzyme tryptophan oxygenase. The enzyme is coded for by the yellowish gene of L.cuprina and vermilion gene of D.melanogaster. Mutants at these loci behave biochemically in a very similar way, both accumulating tryptophan during pupal life (when it cannot be excreted), and both being blocked in brown pigment synthesis; this results in yellow eyed flies in the case of L.cuprina, or red eyed flies in the case of D.melanogaster due to the pteridines. (The pteridine pathway in D.melanogaster is different to the one in L.cuprina and produces the red drosopterin pigments in *D.melanogaster* compared with the yellow sepiapterin pigment in *L.cuprina*) (Summers, 1979). The *vermilion* gene has been cloned (Walker *et al*., 1986a; Searles and Voelker 1986) and the molecular basis of some of its mutants analysed. Experiments designed to clone the *yellowish* gene will be described in Chapter 3.

The second step in the pathway is the removal of the formyl group from N-formylkynurenine to form kynurenine. This reaction is catalysed by the enzyme kynurenine formamidase. In neither *L.cuprina* nor *D.melanogaster* have mutants affecting this enzymatic step been identified. In *D.melanogaster*, two loci code for functional enzymes (Moore and Sullivan, 1978). The probability of mutations occurring simultaneously at both loci is extremely low, which is possibly why mutants at the loci have never been recovered. This might also be the case in *L.cuprina*. The level of activity of kynurenine formamidase in both *D.melanogaster* and *L.cuprina* is very much greater than that of tryptophan oxygenase (Moore and Sullivan, 1978; Summers, 1979), so that conversion of tryptophan to kynurenine is often treated as a single step.

The third step in the pathway is the addition of an hydroxyl group to kynurenine to form 3-hydroxykynurenine. This reaction is catalysed by the enzyme kynurenine 3-hydroxylase, which is coded by the *yellow* gene of *L.cuprina* and the *cinnabar* gene of *D.melanogaster*. *Yellow* and *cinnabar* mutants are very similar biochemically; they accumulate kynurenine during pupal life, and are blocked in xanthommatin production, again resulting in yellow and red eyes respectively.

The *yellow* and *yellowish* genes are non-autonomous, as a dietary supplement of 3-hydroxykynurenine during larval life allows storage of the precursor in the malpighian tubules which provides for some adult xanthommatin production (Summers, 1979).

The final step in the ommochrome pathway is the condensation of two molecules of 3-hydroxykynurenine to form xanthommatin. This step is catalysed by the enzyme phenoxazinone synthase. The *tangarine* locus is thought to be associated with this enzyme, however it is not clear whether it codes for the enzyme itself or for a regulatory protein. *Tangarine* mutants accumulate 3-hydroxykynurenine during pupal life but they have orange eyes indicating that the blockage in xanthommatin production is not complete (Summers, 1979). Two *D.melanogaster* mutants, *cardinal* and *karmoision*, share some similar features with the *tangerine* mutants but it is not clear which one of these genes is homologous to *tangerine*.

There are two additional loci which have been shown to affect brown pigment production in *L.cuprina* and these are *white* and *topaz*. Their exact function has not yet been elucidated, however they are thought to be involved in the uptake, intracellular transport and storage of brown pigment and brown pigment precursors. Unlike *topaz*, the *white* gene affects the pteridine as well as the ommochrome pathway.

Mutations in the *white* gene can either completely abolish brown and yellow pigment production, resulting in white-eyed flies or only reduce the level of pigment production, yielding partially pigmented eyes. *L.cuprina white* mutants behave biochemically in a very similar manner to *D.melanogaster white* mutants (Summers, 1979), strongly arguing that the two genes are homologous. Since the two *white* genes have now been cloned and shown to possess considerable sequence homology, this assumption has been supported by molecular data as well (O'Hare *et al*., 1984; A.J. Howells, pers. comm.).

The *topaz* gene and its mutants are the major topic of this thesis, therefore the biochemical aspects concerning its role in the pathway will be described in more detail. Mutations at the *topaz* locus reduce significantly the level of brown pigment produced. In the two available mutants, *topaz*^{1} and *topaz*^{2}, brown pigment level is 3% and 20% respectively relative to wild type levels (Summers, 1979), so that *topaz*^{1} flies have yellow eyes while *topaz*^{2} eyes are orange.

Examination of the effect that the *topaz* mutations have on the intermediate steps of the pathway show that (i) the mutants fail to accumulate 3-hydroxykynurenine in their malpighian tubules during larval life, and (ii) there is a low level of conversion of kynurenine into 3-hydroxykynurenine in both whole larvae and adult eyes (Summers, 1979; Summers and Howells, 1980). The possible homologous gene to *topaz* in *D.melanogaster* is *scarlet*. *Scarlet* mutants also fail to accumulate 3-hydroxykynurenine in their larval malpighian tubules, however conversion to 3-hydroxykynurenine of the kynurenine that is taken up by the tubules and eyes appears to be normal.

Another major difference between *topaz* and *scarlet* is in the ability of their eyes to take up kynurenine. The uptake and storage of kynurenine in *topaz* eyes appear to be normal (Summers, 1979) while *scarlet* shows a defect in the precursor uptake and storage, reducing its level relative to wild type by about 65% (Sullivan and Sullivan, 1975).

These results suggest that while the *scarlet* mutation affects the uptake mechanisms in both the malpighian tubules and the eyes, the *topaz* mutations affect the malpighian tubules only (Summers 1979). These differences can perhaps be correlated with the altered way that ommochrome pathway operation appears to be coordinated in the two species. As discussed earlier, in developing pupae of *L.cuprina* most of the production of 3-hydroxykynurenine seems to occur in body tissues, (probably in the malpighian tubules), whereas in *D.melanogaster* a much larger emphasis is on its production in the eyes. The significance of the observation that kynurenine conversion to 3-hydroxykynurenine is more defective in topaz mutants than in scarlet is difficult to evaluate. Therefore although it seems likely that the topaz and scarlet genes are homologous, since they are both necessary for the normal uptake and processing of ommochrome precursors, some differences may be encountered in the structures and temporal- and tissue-specific expression of the genes to explain the altered functions.

As outlined earlier, this thesis involves the detailed analysis at the molecular level of the topaz gene region. Consequently a brief review of DNA sequence organization in the eukaryotic genome will now be presented.

1.7 The organization of DNA sequences in the eukaryotic genome

The DNA of the eukaryotic genome can be divided broadly into three major classes: unique, moderately repeated and highly repeated. All eukaryotic genomes appear to contain all three classes of DNA, but there are major variations in the internal organization of these sequences, in particular in the repeated DNA component. Their relative copy number, distribution, size and actual base sequences can differ dramatically between species. The general features of the three classes of DNA sequences are described below.

1.8 Unique sequences

Sequences present in 1 to 10 copies per genome are regarded as unique sequences. Genes are probably the best characterized component of this class (although sequences within genes are not always strictly unique as introns of some genes contain repeated sequences (Karlsson and Nienhuis, 1985) and some genes are repeated, as will be discussed in the next section). Non-genic unique sequences are probably under-represented in reported studies because of the difficulty of relating sequence to obvious function for this type of DNA.

In the structural sense, a gene can be defined loosely as the minimal stretch of DNA which contains the information for its function plus all the sequences required for its correct temporal and spatial expression. The fact that the coding regions of eukaryotic genes are usually not continuous but are interrupted by introns has been well-established over the past 10 years. Thus, more recently, attention has focused on the non-coding regions which lie outside or among the coding regions. The functional significance of the fine scale organization of sequences in these regions has been partially elucidated through comparative studies between genes which have focused on the conservation of sequences outside the coding regions, pointing to possible regulatory signals (Breathnach and Chambon, 1981; Mount, 1982; Keller and Noon, 1985). Recent studies on the analysis of mutations which are caused by insertions, deletions or base substitutions outside the coding regions are also making a significant contribution to our understanding of these regions. In some cases the use of gene transformation systems has allowed the establishment of the actual size of genes (i.e., the length of DNA which needs to be reintroduced to fully restore gene function) and through in-vitro site-specific mutagenesis the exact sequences associated with particular regulatory functions have been identified (Levis et al., 1986).

As might be expected, the emerging pattern is complex, with some sequences being conserved because their actual sequence is functionally important, while others (which still appear to be functionally important) have diverged. In the latter case their role might be explained in terms of the spacing of regulatory signals or the formation of secondary structures. Different genes appear to have different sets of regulatory signals, which can be positioned in various sites along the gene. The common organizational features (for genes transcribed by RNA polymerase II) are as follows:

(1) The 5'-untranscribed region. Since transcription is initiated from the 5' end, the region of the gene preceding transcription initiation site is usually called the 5'-untranscribed region. In most cases it contains a TATA box sequence (TATA$\frac{A}{T}$A$\frac{A}{T}$) which is thought to signal to the RNA polymerase the position of transcription initiation, about 30 bp downstream. Mutations altering the position of the TATA box or its actual sequence reduce the frequency of transcription or cause multiple transcription initiation sites (Breathnach and Chambon 1981). In yeast, examples have been found where transcription initiation appears to be determined by specific DNA sequences rather than simply by the distance from the TATA box, implying a different role for TATA boxes in yeast (Nagawa and Fink, 1985; Hahn et al., 1985). Upstream of the TATA box, at variable distances but usually within a few hundred bp, a CCAAT box is often found; the exact function of this sequence is unknown but it appears to be important in determining the frequency of transcription. Other short sequences (in the order of 10-20 bp) concerned with the tissue specificity or response specificity have also been found, usually within 1 kb of this region (Karin et al., 1984; Pelham, 1982; Walker et al., 1986b).

(2) The transcribed region. This region can be divided into four parts: the 5'-untranslated region, which lies between the transcription

initiation site and the codon for the initiating methionine residue; the exons, which are the protein coding sequences; the introns, the non-coding intervening sequences which are spliced out of the primary transcript; and the 3' untranslated region, which lies beyond the termination codon. The DNA sequence of the exons reflects the amino acid sequence of the protein it codes for, with the existence of an open reading frame (i.e., the absence of termination codons) being perhaps their most general feature.

The presence of introns within most eukaryotic genes studied so far has raised many speculations about their origin and function. They may represent examples of "selfish DNA", conferring little or no phenotypic advantage (Doolittle and Sapienza, 1980) or they may have some function (Gilbert, 1978). Observations supporting both schools of thought have been reported. On the one hand, in different species, related genes with highly conserved coding regions contain introns which have completely diverged in size and sequence (Efstratiadis et al., 1980; Cornish-Bowden, 1982) and this argues for the "selfish DNA" theory. On the other hand, other lines of evidence argue that intron sequences do have a function, even if it is not sequence specific. These are supported by the finding of regulatory sequences within introns (Gillies et al., 1983), the presence of conserved blocks of sequences within introns from functionally related genes (Milner et al., 1984) or the correlation of the position of some introns with the functional domains of proteins (Craik et al., 1981) or with the amino acid residues located at the protein surface (Craik et al., 1982). Structural features generally shared between introns have been identified; at the 5' splice junction, (the donor site) a 9 bp purine rich consensus sequence has been established and at the 3' splice junction (the acceptor site) a 5 bp consensus sequence preceded by a pyrimidine rich tract is found (Breathnach and Chambon 1981; Mount, 1982).

The 3' untranslated region often contains the AATAAA sequence located a few hundred bp downstream from the translation termination codon. This sequence, commonly referred to as "the poly A addition signal" appears now to be primarily involved in endonucleolytic cleavage of the RNA transcript, followed by polyadenylation (Proudfoot, 1984; Montell et al., 1983; Higgs et al., 1983). Other regulatory sequences involved in temporal and tissue specificity have been found downstream from the AATAAA signal (Choi and Engel, 1986).

1.9 Repeated DNA sequences

The presence of repeated DNA sequences in a wide range of eukaryotic genomes was first suggested by reassociation experiments (Britten and Kohne, 1968). Since then many of these repeated sequences have been isolated and submitted to molecular biological investigations revealing features such as size of various repeats, copy number, genomic distribution, nucleotide sequence and degree of sequence conservation of repeats within and between species. These studies divide the repeated sequences into a number of broad categories: (1) repeated gene families, (2) short dispersed repeats, (3) long dispersed repeats, and (4) tandemly arrayed highly repeated sequences (satellite DNA). It should be emphasised that this division is for convenience of description only and both structural and functional overlap between the classes does exist.

1.10 Repeated gene families

Repeated genes can be found as a cluster in one or a few chromosomal locations, or dispersed throughout the genome. Examples of clustered and tandemly arrayed genes are the 18S and 28S ribosomal genes in Xenopus laevis, where about 500 copies of the genes are located in the nucleolus organizer region (Long and David, 1980), or the 160 copies of the 5S RNA genes from D.melangoaster (Hershey et al., 1977). The histone, actin and transfer RNA genes are examples of repeated genes which in different organisms can be found either clustered or dispersed throughout the genome (Kedes, 1979; Fyrberg et al., 1980; Yen and Davidson, 1980 and Clarkson et al., 1973). The presence of multiple copies of a gene may be advantageous to the organism if the gene product (RNA or protein) is required rapidly and in large quantities. An extreme example of the rapid evolution of repeated genes under strong selective pressure is the amplification of dihydrofolate reductase genes in tissue culture cells treated with methotrexate (Kaufman et al., 1979).

Another class of sequences, which belongs somewhere between multigene families and repeated sequences, are the pseudogenes. These sequences, which are often highly homologous to known functional genes, are incapable of producing a functional product due to single or multiple mutations in their regulatory or coding regions. Some are thought to have been generated by gene duplication events, although the exact mechanism is not known. The finding that many pseudogenes lack introns, i.e. resemble a cDNA copy of a processed mRNA, led to the suggestion that they were generated by a reverse transcription of mRNA species followed by integration into the genome. These pseudogenes are now refered to as processed retropseudogenes (reviewed by Little, 1982; Weiner, et al., 1986). The fixation of pseudogenes in the genome, despite their lack of any obvious function, has been interpreted as being the consequence of the inability of the organism to remove such sequences, resulting in general expansion of the genome (Hutchinson et al., 1984).

1.11 Short dispersed repeats

Short interspersed repeated sequences, or SINE (Singer, 1982), refers to repeated sequences, around 300bp long, interspersed in the genome with longer, unique sequences. This pattern of sequence organization, called short periodicity interspersion or the Xenopus pattern is found in most eukaryotic genomes examined so far (for example Davidson et al., 1973; Goldberg et al., 1975; Jelinek and Schmid, 1982 and Bradfield et al., 1985). Reassociation kinetics, used initially to identify this genomic organization, has also pointed to a number of exceptional cases such as D.melanogaster and Apis mellifera whose genomes consist of long repeats (on average 5,600 bp long) interspersed with much longer unique sequences (Manning et al., 1975; Crain et al., 1976a; Crain et al., 1976b); this pattern was called long periodicity interspersion. With the introduction of recombinant DNA techniques to the study of repeated sequences, this distinction in genomic organization patterns was found not to be strictly correct as short repeats have been found in some gene regions from D.melanogaster (Healy, 1985) and long repeats have been found in organisms characterized for their short periodicity interspersion (Jelinek and Schmid, 1982). However in the general sense these terms still describe how the majority of the genome is organized in different species.

One of the best characterized families of the short interspersed repeats are the Alu sequences, found primarily in primate genomes (Schmid and Jelinek, 1982). These sequences, which are approximately 300 bp in length and are present in about 500,000 copies in the human genome, appear to have descended from the 7SL RNA gene and are now known as dimeric retropseudogenes (Ullu and Tschudi, 1984). The observation that most Alu repeats are flanked by direct repeats (Van Arsdell et al., 1981), together with the identification within the Alu family consensus sequence of two regions homologous to promotor sequences for RNA Polymerase III, (Deininger et al., 1981) led to the suggestion that the Alu family is in fact a transposable element moving by self-regulated transcription i.e. via an RNA intermediate (Jagadeeswaran et al., 1981; Rogers, 1983). Sequence studies on other families of SINES indicate that they are mostly retropseudogenes derived from tRNA or tRNA genes (reviewed by Weiner et al., 1986).

1.12 Long dispersed repeats

Long repeats, on average 5-8kb in length, can be found in a wide range of organisms (for example Jelinek and Schmid, 1982; Kay and Dawid, 1983; Sentry and Smyth, 1985). Differences in their structure and repetition frequency further divides them into three broad categories: transposable elements, scrambled repeats and other long repeats.

1.12.1 Transposable elements

Transposable elements, which can be defined as discrete genetic elements capable of transposition in the genome from one chromosomal location to another, have 4 general common properties*: (i) they are present in multiple copies in the genome, (ii) they are found in different

* These criteria for defining transposable elements do not take into account the mounting evidence that other types of repeat elements appear to transpose as well, but will be used in this introduction to simplify matters.

chromosomal locations in different populations of the species, (iii) they possess terminal repeats, and (iv) they create a short duplication of the target site within which they inserted. Further verification of transposition can be obtained either genetically (usually reversion of mutations), by in situ chromosome hybridization (demonstrating different chromosomal locations in different strains) or by direct DNA sequencing of wild type and mutant genes (caused by the insertion of a transposable element).

Transposable elements have been found in a wide variety of organisms such as bacteria, yeast, nematodes, plants and insects (reviewed in Mobile Genetic Elements, edited by Shapiro, 1983; Finnegan, 1985). They have raised particular interest as the demonstration of their transposition was the first indication of fluidity in the genome. In addition, their association with the generation of spontaneous mutations, and more recently their use in DNA transformation experiments (Spradling and Rubin, 1982; Rubin and Spradling, 1982) have encouraged much work to be directed towards the isolation and characterization of transposable elements. While Drosophila species appear to contain 30-50 different families of transposable elements, with D.melanogaster leading the way (Dowsett and Young, 1982; Finnegan, 1985), other organisms, (with the exception of corn, Zea mays,) appear to have only one, or a few families of transposable elements in their genome. Whether this reflects a less advanced state of study relative to Drosophila for other organisms or represents a significant difference in genome structure is still too early to say.

The transposable elements from Drosophila have been divided on the basis of their structure into 3 main groups. So far transposable elements isolated from other organisms fit into these structural groups and will not be described individually.

(1) Copia-like elements. Copia-like elements are named after one of the first transposable elements isolated in D.melanogaster (Finnegan et al., 1978). They consist of internal coding sequences about 5 to 8 kb long, and are flanked by 200-500 bp direct repeats which have short inverted repeats at their termini. There appear to be between 20 to 60 copies of each copia-like element in D.melanogaster embryos, and about double that number in tissue culture cell lines (Potter et al., 1979; Young, 1979). It has been estimated that there are about 30 different types of copia-like elements in the D.melanogaster genome.

Copia sequences have been also found in the form of extra-chromosomal circles (Flavell and Ish-Horowicz, 1981). Their similarity in structure to the proviral form of retroviruses suggests they might be evolutionary related and has lead to the use of the new term "retrotransposons" for these elements. It is as yet unknown what regulates transposition of copia-like sequences however regulatory regions within copia are currently being analysed (Sinclair et al., 1986). Another well characterized member of this family is the Ty element from yeast (Roeder and Fink, 1983).

(11) Fold-back elements. The fold-back, or snap-back elements from D.melanogaster (DmFB) were isolated on the basis of homology to a class of sequences which reassociates at zero C_0t values (Potter et al., 1980; Truett et al., 1981). These elements have inverted terminal repeats and an internal region, both of which can vary in size between different members of the family. The inverted repeat termini are composed of smaller and simpler internal repeats. Towards the distal end 10 bp repeats are interspersed with other DNA sequences; as you move internally, these repeats change progressively to tandem arrays of 21bp, 33bp and finally

155bp repeats. The internal region varies between elements or could be missing altogether (Potter, 1982). Evidence for the mobility of fold-back elements comes from *in situ* hybridization studies which show significant differences in the chromosomal locations of the 30 copies of the element between strains, and from the association of fold-back elements with the $\underline{w}^{Dzl}$ and $\underline{w}^{c}$ unstable mutations (Levis *et al*., 1982; Collins and Rubin, 1983). Similar sequences have also been described in *Strongylocentrotus purpuratus* (Liebermann *et al*., 1983).

(111) *P-elements*. The P-element from *Drosophila* has attracted considerable interest because of its role in hybrid dysgenesis, a syndrome of genetic traits that include sterility, elevated rates of male recombination, mutation and chromosomal aberrations (Bingham *et al*., 1982; Bregliano and Kidwell, 1983), and because of its use as a vector to transform *Drosophila* embryos with DNA sequences of choice (Spradling and Rubin, 1982; Rubin and Spradling 1982). The complete P-element is 2.9kb long, although truncated forms derived from internal deletions have been found. The elements are flanked by short (31bp) inverted terminal repeats which are essential for transposition (O'Hare and Rubin, 1983). There are four open reading frames within the intact P-element, and all four combine to code for an 87,000 dalton protein believed to be a transposase (Karess and Rubin, 1984; Rio *et al*., 1986). Regulation of P-element transposition is determined by differential mRNA splicing, such that functional transposase is made only in the germline (Laski *et al*., 1986). This germline specificity of P-element transposition is in contrast to the Tcl element from the nematode, *Caenorhabditis elegans* where transposition occurs primarily in the somatic tissues (Emmons and Yesner, 1984). Another *Drosophila* transposable element with similar genetic traits to the

P-element is the I-element, also associated with hybrid dysgenesis (Bregliano and Kidwell, 1983). The I-element has been cloned and partially characterized. The element is 5.4 kb in length and does not show structural or sequence homology with any of the other known transposable elements from D.melanogaster (Bucheton et al., 1984).

Features common to all Drosophila transposable elements are the presence of short inverted repeats at their termini, duplication of integration target sites, controlled copy number in the genome (individuals from any particular strain appear to have a fixed characteristic copy number) and a role in the generation of spontaneous mutations (to be described in Chapter 5). Transposition itself is probably determined and regulated differently in the various elements. Copia was shown to be transcriptionally responsive to environmental stress such as heat shock or chemicals (Strand and McDonald, 1985) and transposition bursts of different transposable elements in the course of hybrid dysgenesis have also been reported (Gerasimova et al., 1984a; Gerasimova et al., 1984b). In yeast, lowering the temperature from 30^{0}C (the optimum) can increase the rate of Ty transposition by up to 100 fold (Paquin and Williamson, 1984); this could be due to temperature sensitivity of enzymes involved in transposition. In addition, transcription of the Ty element is induced by ultraviolet light (Rolfe et al., 1986)

1.12.2 Scrambled repeats

The term "scrambled repeats" refers to DNA elements which are composed of a number of different repeated sequences. Each repeat is present in a different copy number in the genome, and in combination with different repeated sequences at any location, making each such cluster a unique combination of sequences. Examples of scrambled repeats have been found in the chicken genome (Eden, 1980; Musti et al., 1981), Drosophila (Wensink et al., 1979) and the slime mould, Physarum polycephalum (Peoples and Hardman, 1983). The fact that each "scrambled cluster" is composed essentially of the same sequences only in different combinations led Wensink et al. (1979) to speculate that these repeats transpose around the genome via a recombination mechanism that has partial site specificity.

1.12.3 Other long dispersed repeats

There have been a number of reports concerned with long dispersed repeats of a different type to the ones already described, for example the 1723 element from X.laevis (Kay and Dawid, 1983) or the Bam family from Lilium henryi (Sentry and Smyth, 1985). Another major repeat family is the LINE or L1 family, found primarily in mammals (Rogers 1983; Singer et al., 1983; Demers et al., 1986). Repeats from the L1 family can be transcribed; they vary in length but are characterized by an A rich stretch of DNA at one end, and by the presence of long open reading frames within their sequence. These features have led to the suggestion that the L1 family are retroposons and have derived from a protein coding gene which spread throughout the genome via RNA intermediates and a reverse transcriptase process, during which most of the copies became non-functional (Hardman, 1986). This hypothesis is supported by the finding of significant sequence homology between several RNA-dependent DNA Polymerases and L1 sequences (Hattori et al., 1986). They speculated that the polymerase binds to its own messenger RNA immediately after translation to start cDNA sythesis, and thus possibly enabling preferential and active dispersion of the L1 repeat family during evolution.

If indeed all dispersed repeated sequences are mobile (in the evolutionary sense), it points to genome fluidity on a scale not perceived only a few years ago, and further emphasizes the need for more work to be done on these sequences before their role in the genome and in evolution can be fully understood.

1.13 Tandemly arrayed highly repeated DNA sequences

Since the discovery of highly repeated DNA sequences by reassociation experiments (Britten and Kohne 1968) a great deal of study has been done on those repeats, characterizing their nucleotide sequences, sequence complexity, genomic organization and chromosomal distribution. These sequences can be repeated thousands of times in the genome and are usually organized in tandemly arrayed blocks. A repeat unit can be as short as only 2 base pairs (bp) long, as for example the ATAT repeat found in the crab genome (Sueoka and Cheng 1962), or can be much longer and more complex (such as the 2500 bp repeat unit of the red necked wallaby satellite, Dennis et al., 1980). In D.melanogaster junction molecules between different highly repeated DNA species have been isolated, indicating that long arrays of repeating sequences can be adjacent in the genome (Peacock et al., 1977).

The chromosomal locations of highly repeated DNA sequences have been determined by in situ chromosome hybridization. The first experiments (Pardue and Gall 1970; Jones 1970), showed that in the mouse the major highly repeated sequences are located in heterochromatin near the centromeres. Heterochromatin (Heitz 1928) does not uncoil at mitotic telophase and is maintained in a condensed state throughout interphase and into the subsequent prophase of the next division cycle. With more data available, it has become clear that there is a good correlation between heterochromatin (often visualised by C-banding) and the location of highly repeated sequences. However, they are not restricted to the centromere and can be present at telomeres and in interstitial bands on chromosomes (Appels et al., 1978; Gerlach and Peacock, 1980).

Examination of highly repeated sequences from different related species shows in many cases that these sequences have been conserved during evolution. The same sequence can be found in species which have been separated for millions of years, and the degree of sequence conservation can be as high as the conservation of some of the structural genes during the time period that the species have separated (Moore et al., 1978).

Salser et al. (1976) proposed an explanation for the presence of the same highly repeated sequence in different species, known as the "library hypothesis". They proposed that each species contains an array or library of different repeat units present at low levels in the genome. Some or all components of the library are held in common by different species. A species produces highly repeated sequences by randomly amplifying some of the repeat units in the library.

Southern (1970) envisaged the generation of highly repeated DNA sequences as involving a rapid initial saltatory replication step in which a stretch of DNA is amplified many times in a tandem fashion, followed by a slow process of divergence by mutations. By combining the two processes and varying the number of amplification steps and the extent of divergence, this model can, in principle, account for the structures of the different highly repeated DNA's as reported in different species.

An alternative theory was proposed by Smith (1976) who states that "DNA whose sequence is not maintained by selection will develop periodicities as a result of random crossover". His theory is based on the assumption that once selection pressure does not apply to conserve a certain sequence, that sequence accumulates random mutations at a high rate. At some stage regions of homology within that sequence are formed, allowing out of register pairing, so consequently unequal crossover would take place causing the elimination or generation of tandemly arrayed repeated sequences. Most of those duplications are eliminated by natural selection but some are fixed by random drift appearing as a monomer repeat unit.

These two proposed mechanisms of amplification of repeated DNA can account for both elimination and generation of repeated DNA during evolution and in most cases the models can not be distinguished on an experimental basis.

1.14 Scope of thesis

Despite the fact that the sheep blowfly, L.cuprina and the fruitfly D.melanogaster diverged about 100 million years ago (Beverly and Wilson 1984), they still retain similar genetic linkage groups (Foster et al., 1981). Biochemical studies on the biosynthesis of eye pigments have shown that the ommochrome pathway operates in a very similar way also in the two species (Summers et al., 1982). The work to be described in this thesis aims to extend to the molecular level knowledge about the genome of the sheep blowfly in general, and about the nature of the genes involved in the ommochrome pathway in particular.

The majority of the work has focused on the topaz eye colour gene; this gene was chosen for a number of reasons. (1) The homologous gene from D.melanogaster, (scarlet) had been cloned and could be used as a probe to isolate the topaz gene. (2) The role of topaz (and scarlet) in the ommochrome pathway was not well understood and their gene products had not been isolated or even identified. It was therefore anticipated that by a detailed comparison of the two genes, the coding regions could be recognized and a putative gene structure and protein sequence determined, which might provide some insight about function. Such a comparison would be of interest from the evolutionary point of view as well. (3) Two spontaneous topaz mutants were available; their analysis could provide information about the sources of natural mutations in L.cuprina and perhaps lead to the isolation of transposable elements from this species. (4) The development of a gene transformation system for the blowfly is an essential step towards assessing the potential role of recombinant DNA technology in the biological control of pest insects. Topaz, being an eye colour gene, could serve as a sensitive phenotypic marker for the development of such a system. (5) A detailed molecular analysis of topaz (or any other gene) would provide information about the general organization of sequences in the L.cuprina genome.

The remainder of the thesis is organized as follows: Chapter 2 provides details about the methods used. In Chapters 3 and 4 the isolation and characterization of the wild-type topaz gene is presented. Questions concerned with the genomic organization of sequences from the topaz gene region, the homology between the topaz and scarlet genes, their structure and the structure of the putative topaz protein are addressed. In Chapters 5 and 6 the isolation and characterization of the two topaz mutants is described. The mutant genes are compared to the wild type gene, and the changes observed between the three alleles analysed. In the final chapter the results obtained in this thesis are discussed with reference to other eye colour genes. The use of topaz and scarlet as markers in gene transformation experiments are described, and the use of such a system for the functional analysis of the topaz mutants considered.

CHAPTER 2

MATERIALS AND METHODS

2.1 CHEMICALS AND REAGENTS

The chemicals and reagents used in experiments are listed in Table 2.1 and enzymes in Table 2.2. Unless otherwise noted the quality of chemicals and reagents used was of analytical grade or higher.

2.2 BACTERIAL, PHAGE AND PLASMID STRAINS

The strains of E.coli, its plasmid vectors and phages λ and M13 which were used in experiments are listed in Table 2.3.

2.3 COMMONLY USED SOLUTIONS

Commonly used media, buffers and other solutions are listed in Table 2.4.

2.4 GROWTH OF BACTERIA, PHAGE AND PLASMIDS

2.4.1 Growth of Bacterial Strains

Bacterial strains LE392 and C600 were propagated in NZCYM whilst JPA101 and JM83 were propagated in LB (Maniatis et al., 1982). Suspensions of the various stocks were kept as 50% glycerol cultures at -20^{0}C. An

∘ Ligations conditions:

Equimolar ratio of vector arms and insert DNA in the presence of T4 DNA ligase.(For ligation buffer see Table 2.4)

Temperature of incubations : 14°C

Incubations time : 18 hours

aliquot of glycerol culture was used to inoculate 20-50 ml of medium in 100-500 ml flasks, and incubated with vigorous shaking at 37^{o}C, usually overnight (o/n), or in the case of preparation of competent cells for plasmid transformations, to an optical density (O.D) at 600 nm of 0.3-0.6. Transformation competent JPA101 were prepared as described by Messing (1983) and JM83 as described by Messing *et al*. (1981). They were either used immediately or stored frozen in aliquots under liquid N_2.

2.4.2 The preparation of genomic DNA libraries

Genomic DNA libraries were prepared by using partially Sau3A digested SWT, *topaz*^{1} and topaz^2 DNA or BamHI digested *topaz*^{2} DNA that were size fractionated on NaCl gradients (16 to 20 kb fractions used) and ligated into the BamHI site of the λ-derived vectors EMBL3A or EMBL4°. The recombinant phage DNA were then packaged into phage particles *in vitro*. Methods for the preparation of vector DNA, for the digestion and ligation reactions and for the preparation of *in vitro* packaging extracts were essentially as described by Maniatis *et al*., 1982. Most of the components were prepared by myself; however, I wish to acknowledge the gift of EMBL4 DNA arms and some samples of packaging extract from Mrs J. Norman, and the assistance of Mandy Walker in packaging the EMBL4 library, both from C.S.I.R.O., Division of Plant Industry. Sonic extracts were prepared by Mrs A. Vacek in our laboratory.

2.4.3 Screening of Recombinant Phage λ Libraries

Screening of recombinant phage libraries was carried out by mixing phage with aliquots (200μl for 140 mm dia. plates, 50μl for 85 mm dia.

plates) of LE392 (EMBL3A) or C600 (EMBL4); o/n cultures of the cells were grown in the presence of 0.2% maltose, then briefly centrifuged (5000g for 5 minutes (min)) and the cells resuspended in 0.5 volumes of 10mM $MgSO_4$. After incubation for 20 min at 37°C to allow phage to adsorb to the cells it was then mixed with 10 ml (large plates) or 3 ml (small plates) 0.7% agarose in NZCYM. The agarose overlay was poured onto plates containing 1.5% agar in NZCYM. The plates were then incubated o/n at 37°C. Optimally, densities of approximately (c.) 10,000 plaques/plate (large plates for initial library screens) or c. 100 plaques/plate (small plates for plaque purification) were achived. The method of Benton and Davis (1977) was used to make replica lifts onto nitrocellulose filters, except that the filters were placed on 3MM paper saturated in 0.5M NaOH/0.5M NaCl and then 0.5M Tris.HCl, pH 7.5/2xSSC, DNA side up (5 min each), for alkali lysis and neutralisation respectively. After fixation at 80°C *in vacuo* for 2-3h the filters were either stored at room temperature (R.T.) or prehybridised immediately (see below).

After hybridisation and autoradiography, plugs of agarose containing positively identified plaques were removed from plates and placed in 1 ml of SM containing 15μl of chloroform. Aliquots of dilutions were then plated out as above and a lift made from the plate with the appropriate titre for the next round of purification.

2.4.4 Plasmid Transformation

Competent bacteria were prepared as described in Maniatis *et al.* (1982). JM83 was used for pUC transformations. These transformations (and M13 in the next section) involved selection for inactivation of the β-galactosidase gene in the plasmid by insertion of foreign DNA, and thus the presence of white recombinant colonies in a background of blue parental colonies when grown on media supplemented with X-Gal. Ligation mixes were made up to 50µl with TE and 20µl of 10xTCM was added. The solution was gently mixed with thawed aliquots of 125µl of cells (about 10^8 competent cells) and the procedure of Maniatis et al. (1982) followed, except that the cells were heat shocked at 37°C for 5 min. Aliquots (100-200µl) of transformation mixes were plated onto LB plates containing 50µg/ml ampicillin and an overlay of 3 ml top agarose (0.7%) containing 35µl 2% X-gal was added. Plates were incubated at 37°C o/n.

2.4.5 Transformation with Phage M13

To M13 ligation mixes, 200µl of competent JPA101 cells (c. 8×10^8 competent cells) was added and the mixture incubated on ice for 40 min, heat shocked at 37°C for 5 min and then placed on ice. 200µl of an o/n culture of JPA101 (for lawn culture), 10µl 200 mM IPTG, 35µl 2% X-Gal and 3 ml top agarose (0.7%) were then added and the solution plated on LB plates. The plates were incubated at 37°C o/n.

2.5 ISOLATION OF NUCLEIC ACIDS

2.5.1 Precipitation of Nucleic Acids

Solutions of nucleic acids were precipitated by addition of either an equal volume 4M NH_4Ac pH 7 or 1/10 volume 3M NaAc pH 5.5, followed by 2-2.5 volumes of ethanol. The ethanol precipitates were stored at -20°C for 10 min to o/n, and then centrifuged for 15-30 min at 4°C (Eppendorf centrifuge) to recover the pellet. The supernatant was then aspirated and the pellet usually washed in 70% ethanol and aspirated dry before being redissolved in H_2O or TE.

2.5.2 Phage λ DNA

DNA was extracted from large or small scale liquid lysates exactly as described in Maniatis et al. (1982); 10^7 pfu of L.cuprina recombinant phage were used per 10 ml of culture.

2.5.3 Phage M13 DNA

Single stranded M13 DNA was isolated essentially by the method of Messing (1983). White plaques and surrounding top agarose were transferred to 13.5 ml tubes containing 1.5 ml 2xYT and incubated at 37°C for 6h with vigorous shaking. The culture was then transferred to Eppendorf tubes, centrifuged for 30 sec and 1 ml of the supernatants transferred to an Eppendorf tube containing 250µl 2.5M NaCl/25% PEG and stored at R.T. for 15 min or 4°C o/n. The PEG precipitates were centrifuged for 5 min, the supernatants aspirated and the pellets resuspended in 200µl NTE. After extraction with 200µl phenol, the aqueous phase was recovered and precipitated with NaAc/ethanol; the pellet was redissolved in 20µl H_2O. The DNA solutions were heated to 65°C for 10 min and then stored at -20°C.

2.5.4 Plasmid DNA

Large and small scale preparations of plasmid DNA were generally carried out by the alkali lysis method (Maniatis et al., 1982) with the following modifications. A loop of transformed colony or glycerol culture was added to 500 ml (large scale) or 5-50 ml (small scale) LB containing 50µg/ml ampicillin and the suspension incubated o/n at 37^{O}C with vigorous shaking. To pellet the cell debris and bacterial DNA following alkali lysis in the large scale method, tubes were centrifuged at 10,000 rpm for 30 min in an SS34 rotor. In the small scale method lysozyme was not used because it did not greatly increase yields of pUC and pUC-derived plasmids.

2.5.5 Genomic DNA

Genomic DNA was extracted from whole adults either by the large scale method of Miklos et al. (1984) or by modification of the small scale method of Coen et al. (1982). For the large scale method about 5 g of adults were ground in liquid N_2 and then homogenised in buffer (10mM Tris. HCl, pH 8.0, 20mM EDTA) and sarcosyl was added to a final concentration of 2%. Cesium chloride and ethidium bromide were added to final concentrations of 1g/ml and 600µg/ml respectively and the solutions centrifuged at 44,000 rpm for about 40h at 20^{O}C (Type 80Ti rotor). The major genomic DNA band was removed with a 19 gauge syringe needle, extracted several times with NaCl saturated isopropanol to remove the ethidium bromide and dialysed extensively against TE. The DNA was precipitated with NaAc/ethanol and the pellet redissolved to a final concentration of 1mg/ml in TE at 4^{O}C. To completely dissolve the DNA, brief incubations at 37^{O} were usually necessary.

For small scale DNA preparations, single heads from adults were homogenised (in Eppendorf tubes) in 100µl cold homogenisation buffer (0.1M Tris, 0.1M EDTA pH9), 100µl 2% SDS was added and the solution incubated at 65oC for 1 h. 8M KAc was added to a final concentration of 0.8M and the mixture incubated on ice for 30 min. After 10 min centrifugation the supernatant was recovered, 100µl isopropanol was added and the solution was incubated at R.T. 10 min. The DNA was pelleted by centrifuging for 10 min, washed twice in 70% ethanol and resuspended in 20µl TE. About 1-2µg DNA were recovered from a single head.

2.6 DIGESTION, ELECTROPHORESIS AND BLOTTING OF DNA

Restriction digestion of between 10 ng and 10µg of DNA was usually carried out in 10-20µl with the appropriate salt and buffer conditions (as recommended by the manufacturers), at 37^{o}C for 1-4 hours with an excess of enzyme. Electrophoresis was carried out in 0.8% agarose gels in TBE or TAE. Samples to be electrophoresed were diluted with a 1/6 volume of 6x Loading buffer, incubated at 60-65^{o}C for 2-5 min and cooled on ice before loading. Gels were electrophoresed at 10-100V at room temperature for 1-18h. Ethidium bromide (EtBr) was either added to the gel to a final concentration of 0.5µg/ml or the gel was gently agitated in a solution of 0.5µg/ml ethidium bromide for 15-20 min after electrophoresis.

Transfer of DNA from agarose gels to nitrocellulose membrane was carried out essentially as described by Southern (1975).

2.7 PREPARATION OF RADIOACTIVE NUCLEIC ACIDS

2.7.1 Nick Translation

Double stranded DNA (plasmid, phage or genomic) was nick translated essentially as described by Rigby et. al. (1977) with the following modifications (K.C. Reed, pers. comm.). Approximately 200 ng of

DNA was incubated with 100 pg DNAase I in 1 x NTB at 14^{o}C for 15-60 min (incubation time depending on the type and purity of the DNA). The following reagents were added and the solution adjusted to 1 x NTB: dGTP, dATP and dTTP (each to 25μM); 25μCi [α-^{32}P] dCTP and 5U DNA Polymerase I (holoenzyme). The solution was incubated at 14^{o}C for a further 30 min. ([α-^{32}P] dATP was also used in place of [α-^{32}P] dCTP with the appropriate changes in the unlabelled nucleotides).

The reaction was stopped by addition of SDS to 1% and EDTA to 25mM, and unincorporated nucleotides were removed by NH_4Ac/ethanol precipitation (40μg carrier DNA was routinely added to facilitate precipitation and visualisation of the pellet). The pellet was redissolved in a solution containing 1-5mg carrier DNA, and stored at -20^{o}C if not used immediately. Prior to addition to the hybridization solution the probe was boiled for 3 min and then quick cooled. Incorporation was usually 70-90%, giving probes with specific activities of approx. 10^{8} dpm/μg DNA.

2.7.2 M13 Probes

M13 probes were prepared as described for the C-track reaction (section 2.10) only no dideoxy nucleotides were used and the reaction was not chased with unlabelled nucleotides. Following the incubation at 50^{o}C, the DNA was ethanol precipitated using NaAc/ethanol in the presence of carrier DNA (500μg). Prior to addition to the hybridization solution, the probe was boiled for 3 min and then quick cooled. Incorporation was usually 50%, giving probes with 2 x 10^{7}cpm/100ng template used.

2.7.3 Analysis of Incorporation of Label

Incorporation of labelled nucleotides for all methods was assayed by PEI-cellulose thin-layer chromatography. 0.2μl of the incubation mix (both before and after incubation) was spotted onto a 7cm long sheet of PEI-cellulose and the sheet chromatographed in 0.75M KH_2PO_4, pH 3.5, till the front was 1 cm from the top. The sheet was then dried and autoradiographed for 30 min at room temperature. Label incorporated into DNA remains at the origin, while that in unincorporated nucleotides moves just behind the solvent front, allowing a ready estimation of the efficiency of incorporation to be made. Small aliquots of the reaction products were also analysed in a liquid scintillation counter to determine the exact level of incorporation.

2.8 HYBRIDIZATION AND AUTORADIOGRAPHY OF DNA FILTERS

All filters were prehybridized in plastic bags in 50 mM HEPES pH 7.0, 3 x SSC, 0.1% SDS, 0.2% Ficoll, 0.2% BSA, 0.2% PVP, 1 mM EDTA, 100μg/ml carrier DNA at 60°C (high stringency) or 50°C (low stringency) for 16-24 h. A probe (either nick translated or M13) was added to the prehybridization mix and incubated for another 16-24 h. The filters were shaken throughout the prehybridization and hybridization incubations.

Following hybridization, filters were rinsed in 2 x SSC and then washed twice for 30-60 min in 2xSSC/0.1% SDS, 65°C (high stringency) or 50°C (low stringency). After rinsing in 2xSSC the filters were blotted with 3MM paper to remove excess moisture, wrapped in Gladwrap and exposed to Fuji X-ray film in cassettes with intensifying screens at -45°C from o/n to 16 days.

2.9 SUBCLONING

All subcloning into plasmids utilised the pUC series of vectors. The same methods were used to ligate restriction fragments into M13 vectors for sequencing. 1μg of plasmid or M13 DNA was digested with the appropriate restriction enzyme(s), extracted successively with phenol, phenol/IAC and IAC, and then precipitated with NaAc/ethanol. The pellet was redissolved in H_2O at 10ng/μl.

The DNA to be subcloned was digested and sometimes simply treated as above ("shotgun" subcloning of fragments from recombinant phage into plasmid vectors, and from phage and plasmids into M13) or the required fragment was purified. This was accomplished by electrophoresing digests in agarose gels (0.8%), excising the appropriate band from the gel and eluting it from the agarose by electrophoresis in TBE at 100V for 1 h; the gel fragment was contained in a small piece of dialysis tubing containing 400ul TBE. The DNA (which migrates into the TBE within the tubing) was then precipitated with NaAc/ethanol and the pellet redissolved in H_2O.

For ligations, from 10- 100 ng of insert DNA and 10 ng vector DNA were incubated with 1U T4 ligase in 1xLigation buffer at 14°C o/n. Aliquots of ligation mix were then used to transform competent cells.

2.10 DNA SEQUENCING

A modification of the dideoxy of Sanger _et al_. (1978) was used to sequence DNA fragments.

2.10.1 C-tracking

Putative recombinant M13 clones were checked for quality and orientation of inserts by C-tracking. 1µl of redissolved single stranded template (c. 100 ng) was annealed with 2.5µl primer mix (1.5µl 10 x M13 reaction buffer, 1.5µl 70mM $MgCl_2$, 2.5 ng 17-mer sequencing primer, 20µCi [α-^{32}P] dCTP in 11 µl) at 55°C for 5 min. The primer/template mix was then incubated with 2µl dideoxy/Klenow mix (8µl ddC mix, 1U DNA Polymerase I (Klenow fragment), per 4 templates) at 50°C for 15 min and then incubated with 1µl 1mM dNTPs (chase mix) at 50°C for 15 min. 6µl of sequencing dye was then added, the samples boiled for 1 min, quick cooled and 3µl aliquots electrophoresed on 8% acrylamide/urea sequencing gels (acrylamide: bis-acrylamide 40:1, polymerised by addition of APS and TEMED to 0.1%) at 1800V for 60-75 min. Following electrophoresis the gel was transferred to 3MM paper and then vacuum dried at 80°C for 20 min before exposure to Fuji X-ray film in cassettes without screens at room temperature o/n. The C-tracks were then examined and those clones which appeared to have inserts were selected for full sequencing.

2.10.2 Full Sequencing

Full sequencing was carried out essentially as above. 3µl of template was incubated with 11µl primer mix for 5 min at 55°C, cooled to R.T. and 0.5U Klenow fragment added. Four aliquots of 3.2µl were then incubated with 2µl of ddGTP/dGTP, ddATP/dATP, ddTTP/dATP or ddCTP mixes respectively (50°C for 15 min). Again 1µl chase mix was added and incubations were continued at 50°C for 15 min; 6µl dye was then added, the solutions boiled, cooled on ice and 3µl samples loaded onto gels. Both short (40cm) and long (80 cm) 8% sequencing gels were run, at 1800V for 60-75 min and 2700V for 12 h respectively. Gels were dried and exposed as discussed above.

2.11 COMPUTER ANALYSIS

DNA sequences were analysed using the program "The DNA Inspector II" version 1.05µ by Bob Gross from Textco.

TABLE 2.1: CHEMICALS, REAGENTS AND THEIR SOURCES

Chemical or Reagent	Source[a]
Agar	Difco
Agarose	Pharmacia, Sigma
Ammonium persulphate (APS)	Sigma
Ampicillin	Sigma
Acrylamide	BioRad
Bactotryptone	Difco
Bovine serum albumin, V (BSA)	Sigma, BRL
5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal)	Sigma, Boehringer
Casamino acids	Difco
Cesium chloride	Metallgesellenschaft
DNA (salmon sperm)	Sigma
DNA (λ)	Boehringer
DNA (gel markers)	Pharmacia, NEB, Boehringer
Deoxyribonucleoside triphosphates	NEB, Pharmacia
[α-^{32}P]-dATP, [α-^{32}P]-dCTP (1500Ci/mmol)	BRESA
[α-^{32}P]-dCTP (3000Ci/mmol)	Amersham
Dithiothreitol (DTT)	Calbiochem
Ethidium bromide	Sigma
Ethylenediaminetetraacetic acid (EDTA)	Fluka
Ficoll	Pharmacia

Table 2.1 Continued

Formamide	Fluka
Isopropyl-β-D-thiogalactopyranoside (IPTG)	Sigma, Boehringer
M13 sequencing primer	NEB, Pharmacia
2-mercaptoethanol	Aldrich
N,N'-methylene-bis-acrylamide (bis)	BioRad
Nitrocellulose BA85 (0.45μm)	Schleicher & Schuell
NZamine (casein amino acid hydrolysate)	Sigma
Polyethyleneimine (PEI) - cellulose	Merck
Phenol	Wako
Polyethylene glycol (PEG)	Sigma, Koch-Light
Polyvinylpyrollidine (PVP)	Sigma
Sodium dodecyl sulphate (SDS)	BioRad, Ajax
Tetracycline	Sigma
N,N,N',N'-tetramethylethylenediamine (TEMED)	BioRad
Trizma base	Sigma
Urea	BRL
X-ray Film (RX)	Fuji
Yeast extract	Difco

[a]Abbreviated names of manufacturers:

BRL:Bethesda Research Laboratories

BRESA: Biotechnology Research Enterprises S.A.

NEB: New England Biolabs

TABLE 2.2: ENZYMES AND THEIR SOURCES

Enzyme	Source
Alkaline phosphatase (CIP)	Boehringer
DNA Polymerase I(*E.coli*) - holoenzyme	Pharmacia
DNA Polymerase I (*E.coli*) - Klenow	Boehringer, BRESA
DNAase 1	Sigma, NEB
Lysozyme	Sigma
Proteinase K	Boehringer
Restriction endonucleases	Pharmacia, Boehringer, NEB
Ribonuclease A	Sigma
T4 DNA ligase	Pharmacia, NEB

	GENOTYPE
LE392	F^-, hsdR514(r_k^-, m_k^-) supE44, supF58, lacY1 or Δ(lacIZY)6, galK2, galT22, metB1, trpR55, λ^-
C600	F^-,thi-1, thr-1, leuB6, lacY1, tonA21, supE44, λ^-
JPA101	Δ(lac, pro), thi, supE44, recA, Tc^R(::Tn10), ton, lambda^r, [F'traD36, proAB$^+$, lacI^q, lacZΔM15].
JM83	ara, A(lac-proAB), rps1,(strA), Ø80, lacZAM15

TABLE 2.3: BACTERIAL, PHAGE AND PLASMID STRAINS

Strain	Reference/Comments
(a) Bacterial Strains	
LE392	Maniatis *et al.*, 1982
C600	Maniatis *et al.*, 1982
JPA101	This strain is a derivative of JM101 and contains a tetracycline resistance marker. It was constructed by J.P. Adelman (Genentech)
JM83	This strain is also a derivative of JM101, only the β-galactoside gene is constitutive and therefore does not require IPTG for induction. It was received from E. Keshet (pers. comm.)
(b) Bacteriophage Vectors	
λ derivatives	
EMBL3A, EMBL4	Frischauf *et al.*, 1983
M13 derivatives	Messing, 1983
mp8, mp18, mp19	
(c) Plasmid Vectors	
pUC13, pUC19	Vieira and Messing, 1982; Messing, 1983

TABLE 2.4: COMMONLY USED SOLUTIONS

(a) Media

NZCYM:

NZamine	1.0%
NaCl	0.5%
Casamino acids	0.1%
Bacto-yeast extract	0.5%
$MgCl_2$	0.2%
pH adjusted to 7.5	

LB:

Bactotryptone	1.0%
Bacto-yeast extract	0.5%
NaCl	0.5%

2xYT:

Bactotryptone	1.6%
Bacto-yeast extract	1.0%
NaCl	0.5%

SM:

NaCl	100mM
Mg_2SO_4	8mM
Tris.HCl (pH 7.5)	50mM
Gelatin	0.001%

Table 2.4 Continued

(b) Buffers

TE:	
Tris.HCl (pH 7.6)	10mM
EDTA	1mM
NTE:	
Tris.HCl (pH 7.6)	10mM
EDTA	1mM
NaCl	100mM
20xSSC:	
NaCl	3M
Na citrate	300mM
pH adjusted to 7.0	
10x Restriction buffers:	
Tris.HCl (pH 7.5)	500mM
$MgCl_2$	80mM
DDT	1mM
NaCl	according to manufacturers specifications for different enzymes

Table 2.4 Continued

1xTBE:	
Tris-borate	89mM
Boric acid	89mM
EDTA	1mM
1 x TAE:	
Tris-acetate	40mM
EDTA	2mM
6xDNA gel loading buffer:	
Glycerol	30%
Bromophenol blue	0.05%
Xylene cyanol	0.05%
6x Sequencing gel loading buffer:	
Formamide	90%
Glycerol	9%
EDTA	5mM
Bromophenol blue	0.05%
Xylene cyanol	0.05%

Table 2.4 Continued

10x Nick translation buffer:	
Tris.HCl (pH 7.5)	500mM
$MgCl_2$	50mM
BSA	100µg/ml
10 x TCM (transformation buffer):	
Tris.HCl (pH 7.5)	100mM
$CaCl_2$	100mM
$MgCl_2$	100mM
10x Ligation buffer:	
Tris.HCl (pH 7.4)	500mM
$MgCl_2$	100mM
DTT	100mM
ATP	10mM
BSA	1mg/ml
Spermidine	10mM
10x M13 reaction buffer:	
Tris.HCl (pH 7.6)	70mM
EDTA	1mM
2-mercaptoethanol	50mM

Table 2.4 Continued

(c) Other

Carrier DNA:	
Salmon sperm DNA	10mg/ml
The DNA was sonicated to an average length of 300bp.	
ddCTP mix:	
ddCTP	50μM
dATP	100μM
dGTP	100μM
dTTP	100μM
ddGTP mix:	
ddGTP	100μM
dGTP	10μM
dATP	100μM
dTTP	100μM

Dideoxy A and T mixes made up similarly. Precise ddNTP:dNTP ratio for all four mixes was adjusted empirically with each batch to give the appropriate termination pattern.

CHAPTER 3

THE ISOLATION AND CHARACTERIZATION OF THE WILD TYPE *TOPAZ* GENE

3.1 INTRODUCTION

Over twenty loci which affect xanthommatin production have been identified in *D.melanogaster* (Lindsley and Grell, 1968; Phillips and Forrest, 1980) and mutants at these loci, recognized by their obvious phenotypic change in eye colour, were isolated and mapped both genetically and cytogenetically. However, our biochemical understanding of the defects in these mutants is still very incomplete as the gene products of most of them (including *white* and *scarlet*) have not yet been identified. As a result, conventional cloning approaches such as antibody screening or the use of oligonucleotide probes, which require information about the protein, are not applicable. In addition, the ommochrome biosynthetic pathway is a minor one, operating in only a limited number of cells in the whole organism; even those genes that code for known enzymes of the pathway are likely to be expressed at very low levels, so cloning methods based on differential hybridization have been unsuccessful (Evans 1981; R.G. Tearle, pers. comm.). At the present time, three genes involved in the brown pigment pathway have been cloned: these are *white*, *vermilion* and *scarlet*; it is relevant to consider the cloning strategies adopted in their isolation.

The *white* gene was cloned by three independent groups using genetic approaches. All three made use in their cloning strategy of the observation that the *white-apricot* ($\underline{w}^a$) mutation is associated with an insertion of the transposable element *copia* in the *white* gene region

(Gehring and Paro, 1980). Bingham *et al*. (1981) screened a genomic library made from $\underline{w}^a$ DNA with a probe made from *copia* sequences and tested the chromosomal locations of the positive clones by *in situ* hybridization to *D.simulans* polytene chromosomes. (*D.simulans*, a sibling species of *D.melanogaster*, shares the same linkage groups and has similar polytene chromosomes, but has a much lower copy number of *copia* sequences). A recombinant phage hybridizing to region 3C on the X chromosome, which is the cytogenetically mapped location of *white*, was isolated and the regions corresponding to the *white* gene identified by the position of the *copia* insertion. (It was assumed that the insertion was within the gene itself, causing the mutation by disrupting gene function - an assumption which later proved to be correct). Goldberg *et al*. (1982) "walked" along a large transposable element (TE) known genetically to carry the *white* locus. They defined the *white*$^-$ locus first by locating genetically mapped *white* rearrangement breakpoints in their clones and then by using fragments from the breakpoint region to study DNA from strains carrying the $\underline{w}^a$ allele for the location of the *copia* insertion. Pirrotta *et al*. (1983) microdissected the chromosomal bands corresponding to 3C of a $\underline{w}^a$ mutant, made a mini-library from the DNA isolated from this region, and identified clones carrying *white* by the presence of *copia* insertion.

A second gene involved in the brown pigment pathway, *vermilion*, which codes for the enzyme tryptophan oxygenase, was cloned by two different approaches. Walker *et al*. (1986a) used an approach based on the availability of a cloned rat tryptophan oxygenase gene. The rat tryptophan oxygenase cDNA clone was used as a probe to screen a *D.melanogaster* genomic DNA library. Positively-hybridizing clones were then confirmed to carry the *vermilion* gene by *in situ* hybridization and by Southern and Northern blot analysis, using DNA and RNA from wild type and various *vermilion* mutant strains. Searles and Voelker (1986) used a genetic approach for the cloning of *vermilion*. In a P-element mutagenesis experiment a *vermilion* mutation was induced. Genomic DNA libraries were made from DNA of flies carrying this mutation, and screened with a probe made from P-element sequences. Positively-hybridizing clones were sorted by *in situ* hybridization, with the ones hybridizing to band 10A1-2 (the cytological location of *vermilion*) being studied further by Southern and Northern blot analysis.

The cloning of the third gene of the brown pigment pathway, *scarlet*, (Tearle *et al*., manuscript in preparation; Tearle, 1986) depended on genetic studies and was in fact a preliminary step of a strategy aimed at cloning the gene *transformer*, which maps close to *scarlet*. Since the work has not yet been published and the *scarlet* clone served as the basis for the work described in this thesis, the cloning strategy will be described in detail.

The isolation of the *scarlet* gene involved firstly an X-ray mutagenesis experiment designed to induce and recover new *scarlet* mutants. X-ray irradiated males were mated to females carrying a deletion of the *scarlet* region. Progeny flies with the bright red eye colour characteristic of *scarlet* were selected and their progeny examined for chromosomal rearrangements involving the 73A region, the mapped location of *scarlet*, and re-tested for complementation with two other *scarlet* alleles. One *scarlet* inversion was found to break within both the *scarlet* and the *rosy* gene regions. By using cloned fragments from the *rosy* region (Bender *et al*., 1983) to probe Southern blots of genomic DNA prepared from the inversion and parental strain, the exact position of the breakpoint was identified, providing a probe to enter the *scarlet* region via the inversion. This probe was used to screen a genomic DNA library prepared from the inversion stock, and phage containing the rearrangement region were isolated. The DNA fragments in these phage that did not contain known *rosy* region sequences were assumed to come from the *scarlet* region. They were used to probe a wild type genomic library, and phage containing the *scarlet* region were recovered. (This was verified by *in situ* hybridization to polytene chromosomes). The restriction map of the *scarlet* region is shown in Figure 3.1. The inversion breakpoint occured within the 0.8 BamH1/HindIII fragment (map coordinates -0.8 to 0.0) indicating that it is probably part of the *scarlet* gene itself. This has been confirmed by Southern and Northern blot analysis of DNA from wild type and mutants and by DNA sequencing (Tearle 1986; and see Chapter 4).

As outlined in the introduction (Section 1.6) the biochemical studies of Summers and Howells (1978, 1980) suggested that the *white*, *topaz* and *yellowish* genes of *L.cuprina* are homologous to the *white*, *scarlet* and *vermilion* genes, respectively, of *D.melanogaster*. The *white* gene from *L.cuprina* has been cloned on the basis of its homology to sequences in the *white* gene of *D.melanogaster* (A.J. Howells, pers. comm.). The simplest and most readily applicable strategy for the cloning of *topaz* and *yellowish* would be to test whether sequences from the cloned *scarlet* and *vermilion* genes could be used as probes to detect sequences of their homologous genes in *L.cuprina*. This chapter describes these experiments.

3.2 RESULTS

3.2.1 The isolation of clones carrying the *topaz* gene

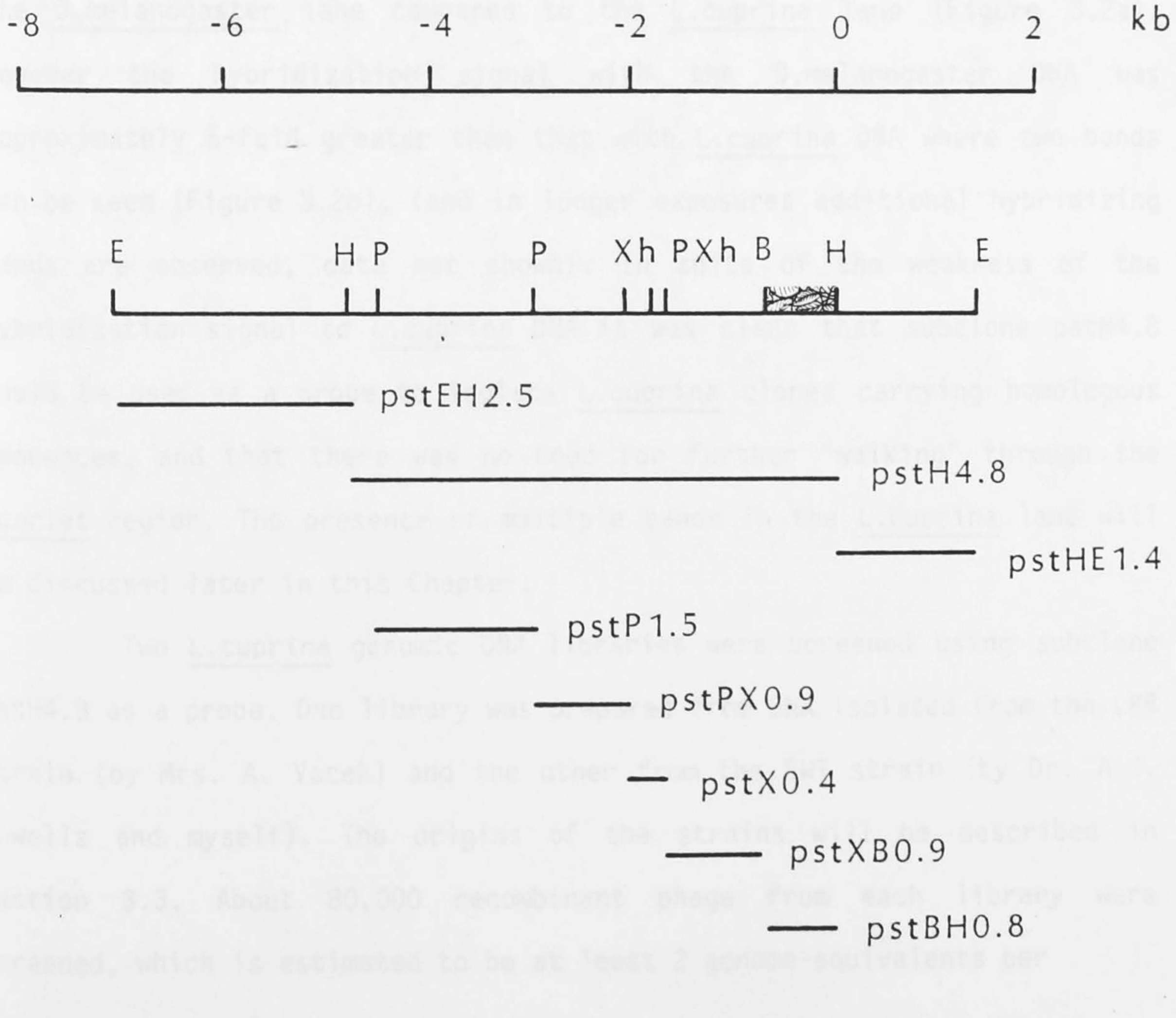
The initial step towards the isolation of the *topaz* gene was to examine the homology, at the DNA level, between the cloned *scarlet* region and sequences in the *L.cuprina* genome. At the time that this work was

FIGURE 3.1 Restriction map of the scarlet gene region.

Restriction sites are: E, EcoR1; H, HindIII; P, PstI; Xh, XhoI; B, BamH1.

The major subclones used in this work are also shown.

▨ Position of the inversion breakpoint.



carried out there was no information regarding the location of the scarlet coding sequences within the cloned scarlet region. It was expected that if scarlet and topaz did share common sequences, they would be coding sequences; therefore it was seen as being a matter of "walking" through the scarlet region to test for homology. As a starting point, a subclone which carries the 4.8 kb HindIII fragment, pstH4.8 (Figure 3.1, -4.8 to 0.0), was used to probe a Southern blot of genomic DNA from the SWT strain of L.cuprina (Figure 3.2). This subclone was chosen as it contains the breakpoint in the inversion strain used in the cloning of the scarlet region (which has a st^- phenotype); therefore it seemed likely that it would contain coding sequences. About half the amount of DNA was loaded in the D.melanogaster lane compared to the L.cuprina lane (Figure 3.2a), however the hybridization signal with the D.melanogaster DNA was approximately 5-fold greater than that with L.cuprina DNA where two bands can be seen (Figure 3.2b), (and in longer exposures additional hybridizing bands are observed, data not shown). In spite of the weakness of the hybridization signal to L.cuprina DNA it was clear that subclone pstH4.8 could be used as a probe to isolate L.cuprina clones carrying homologous sequences, and that there was no need for further "walking" through the scarlet region. The presence of multiple bands in the L.cuprina lane will be discussed later in this Chapter.

Two L.cuprina genomic DNA libraries were screened using subclone pstH4.8 as a probe. One library was prepared from DNA isolated from the LBB strain (by Mrs. A. Vacek) and the other from the SWT strain (by Dr. A.J. Howells and myself). The origins of the strains will be described in Section 3.3. About 80,000 recombinant phage from each library were screened, which is estimated to be at least 2 genome-equivalents per

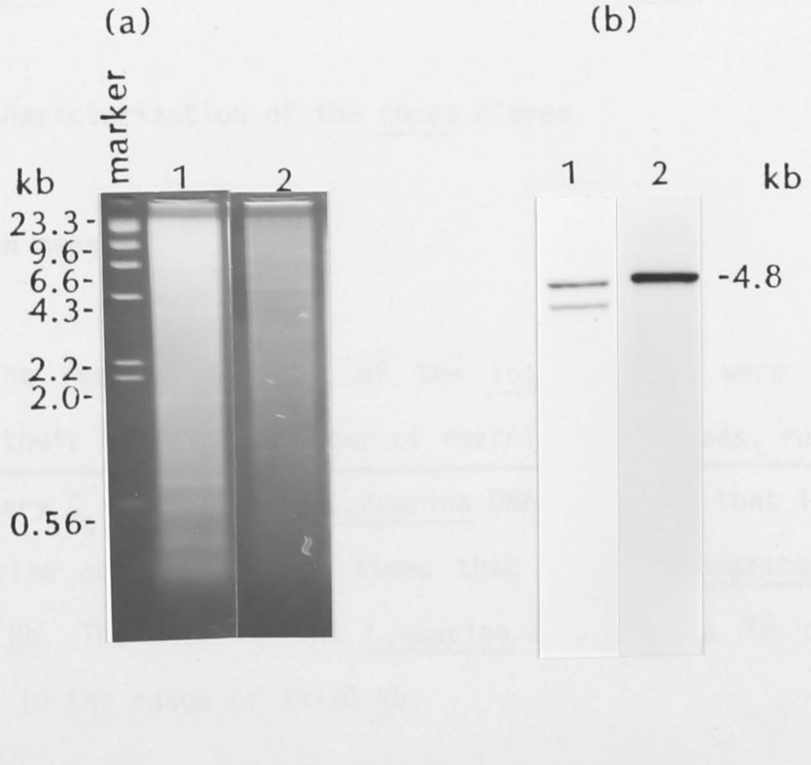
FIGURE 3.2 Hybridization of sequences from the scarlet gene of D.melanogaster to genomic DNA of L.cuprina.

(a) EtBr-stained gel of genomic DNA from L.cuprina (lane 1) and D.melanogaster (lane 2) both digested with the restriction enzyme HindIII.

Marker DNA: Lambda DNA digested with HindIII (to be refered to as λ/HindIII).

(b) Autoradiogram of a Southern blot of gel (a), probed with subclone pstH4.8.

(Nick-translated probe; low stringency hybridization).



library*. Much variation was observed in the plaque size of the initial library platings; some plaques were large while others were so small (pin-prick size) that they were nearly invisible. This variation was particularly noticeable when compared to the relative uniformity in the plaque size of a D.melanogaster genomic DNA library prepared in the same cloning vector. A possible explanation for this will be presented later.

Altogether seven positive clones were obtained from these screenings of the libraries. They were rescreened through three rounds of plaque purification. All seven "positive" clones had very small plaque size. Being slow growers meant that these phage tended to be under-represented relative to other phage during the rounds of plaque purification. Once purified, phage from each clone were amplified and their DNA prepared.

The following work is aimed at showing that these λ clones, selected on the basis of their homology to sequences from the scarlet region, do indeed carry the topaz gene. For simplicity I will refer to them as "the topaz clones" rather than "the putative topaz clones".

3.2.2 Characterization of the topaz clones

Restriction maps

The restriction maps of the topaz clones were established by digesting their DNA with a number of restriction enzymes, running the cut

* Preliminary C_0t analysis of L.cuprina DNA indicates that it has a haploid genome size approximately 4 times that of D.melanogaster, i.e., about 680,000 kb. The size of the L.cuprina DNA inserts in the recombinant phage is in the range of 14-20 kb.

DNA on an agarose gel and analysing the sizes of the fragments obtained; the maps are given in Figure 3.3. Six of the seven clones overlap each other almost perfectly. There is one fragment in λtol that is slightly shorter than its equivalent fragment in the other phage (a 2.8 kb EcoRl Hind III fragment (-3.6 to -0.8) that is 3.2 kb in the other clones). One phage, λtol2, has a restriction map that in part overlaps the other maps but in part is different. The nature of the changes in λtol2 will be considered in the Discussion.

Homology to the *scarlet* region

The regions of sequence homology between the *scarlet* and *topaz* clones were determined by hybridizing restriction fragments from the *scarlet* region to Southern blots of cloned *topaz* DNA, after digestion with various restriction enzymes (e.g. Figure 3.4a). A composite diagram of the regions of shared homology between *scarlet* and *topaz* is shown in Figure 3.4b.

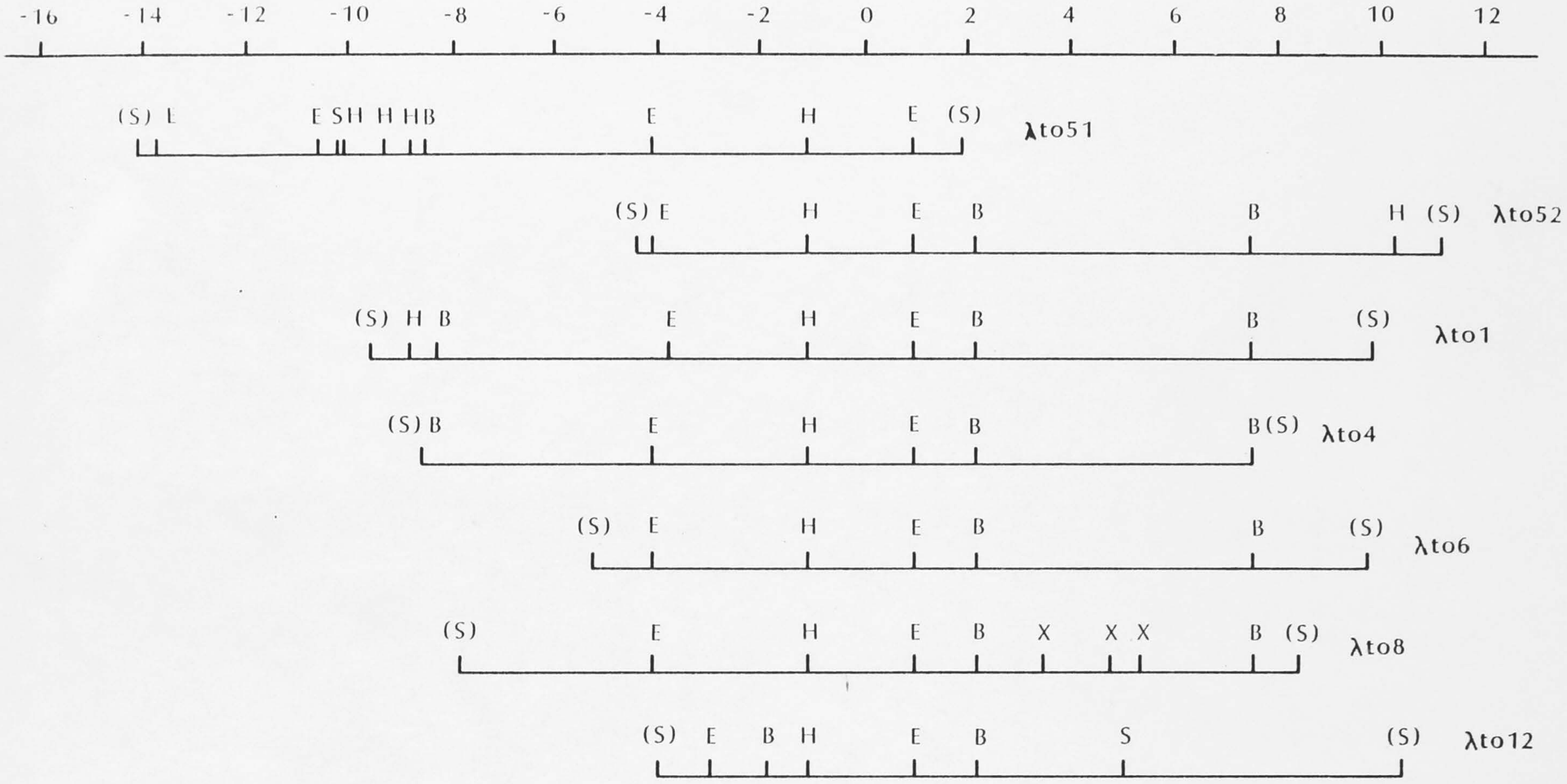
The homology between *scarlet* and *topaz* is co-linear, and covers a region of 6.2 kb in *scarlet* (co-ordinates -4.8 to 1.4) and 6.3 kb in *topaz* (-0.8 to 5.1). When fragments to the left of HindIII (-4.8) or to the right of EcoRl (1.4) of *scarlet* were used as probes, no homology to the *topaz* clones could be detected even with lower hybridization stringency (3xSSC at 50^{o}C) and long exposures of the film.

Organization in the genome

The genomic organization of cloned DNA sequences can be studied by hybridizing these sequences to Southern blots of the genomic DNA. When

FIGURE 3.3 Restriction maps of the topaz clones.

Clones λto51, λto52 and λto12 were isolated from the LBB library. The rest were isolated from the SWT library. All clones are carried in the vector EMBL3A; the SalI sites, (S) at the ends of all inserts come from the poly-linker of the vector. Restriction enzymes used: E, EcoR1; H, HindIII; B, BamH1; S, SalI; X, XbaI (XbaI sites were determined only for the +2.4 to +7.8 region in clone λto8).



-16 -14 -12 -10 -8 -6 -4 -2 0 2 4 6 8 10 12
(S) E E S H H H B E H E (S)
λto51
(S) E H E B B H (S)
λto52
(S) H B E H E B B (S)
λto1
(S) B E H E B B (S)
λto4
(S) E H E B B (S)
λto6
(S) E H E B X X X B (S)
λto8
(S) E B H E B S (S)
λto12

FIGURE 3.4 Homology between scarlet sequences and the topaz clones.

(a) Southern blot analysis (example)

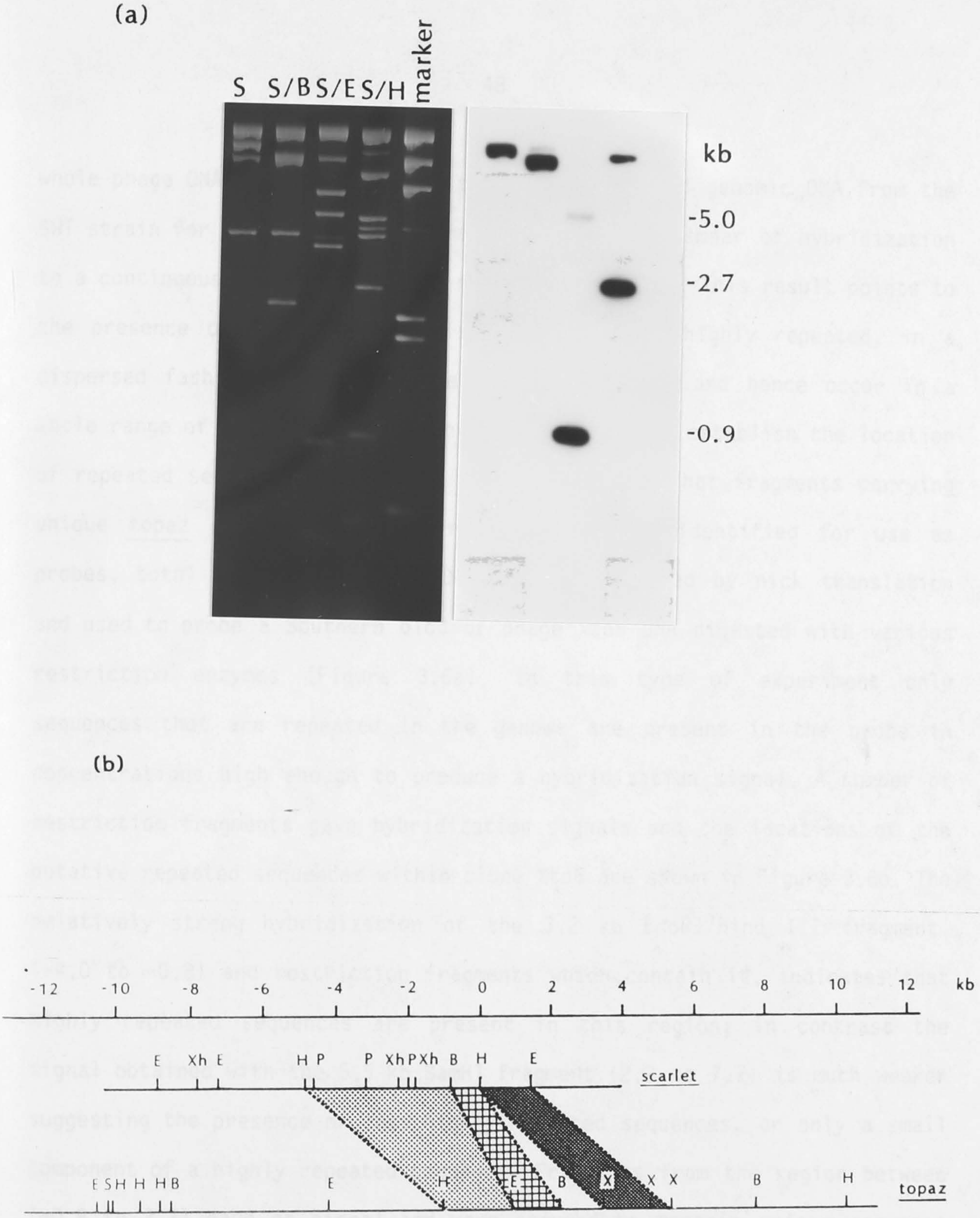
Left: EtBr stained gel. DNA from λto51 digested with the following enzymes S, SalI; S/B, SalI+BamHl. S/E, SalI+EcoRl; S/H, SalI+HindIII.

Marker - λ/HindIII.

Right: Autoradiogram of a Southern blot of this gel probed with subclone pstH4.8.

(Nick translated probe; low stringency hybridization).

(b) The regions of homology between scarlet and topaz.

(a)
S
S/B
S/E
S/H
marker
kb
-5.0
-2.7
-0.9
(b)
-12
-10
-8
-6
-4
-2
0
2
4
6
8
10
12
kb
E
Xh
E
H
P
P
Xh
P
Xh
B
H
E
scarlet
E
S
H
H
H
B
E
H
E
B
X
X
X
B
H
topaz

whole phage DNA of λto51 was used to probe restricted genomic DNA from the SWT strain for such a study, the result was a black smear of hybridization to a continuous range of fragment sizes (Figure 3.5). This result points to the presence of sequences in this clone that are highly repeated, in a dispersed fashion, throughout the L.cuprina genome and hence occur in a whole range of restriction fragment size classes. To establish the location of repeated sequences within the topaz clones, so that fragments carrying unique topaz sequences without repeats could be identified for use as probes, total L.cuprina genomic DNA was ^{32}P-labelled by nick translation and used to probe a Southern blot of phage λto6 DNA digested with various restriction enzymes (Figure 3.6a). In this type of experiment only sequences that are repeated in the genome are present in the probe in concentrations high enough to produce a hybridization signal. A number of restriction fragments gave hybridization signals and the locations of the putative repeated sequences within clone λto6 are shown in Figure 3.6b. The relatively strong hybridization of the 3.2 kb EcoR1/Hind III fragment (-4.0 to -0.8) and restriction fragments which contain it, indicates that highly repeated sequences are present in this region; in contrast the signal obtained with the 5.5 kb BamH1 fragment (2.2 to 7.7) is much weaker suggesting the presence of less highly repeated sequences, or only a small component of a highly repeated sequence. Fragments from the region between (-0.8 to 2.1) gave no signal and were assumed to contain unique sequence DNA.

3.2.3 Further characterization of the topaz λ-clones: subcloning

The presence of repeated sequences within the topaz λ clones meant that before any further analysis of the region homologous to scarlet, or of

FIGURE 3.5 Hybridization of DNA from λto51 to genomic DNA from SWT.

(a) EtBr stained gel. Genomic DNA digested with HindIII (Lane 1) and HaeIII (Lane 2).

(b) Autoradiogram of a Southern blot of gel (a) probed with clone λto51.

(Nick-translated probe; high stringency hybridization).

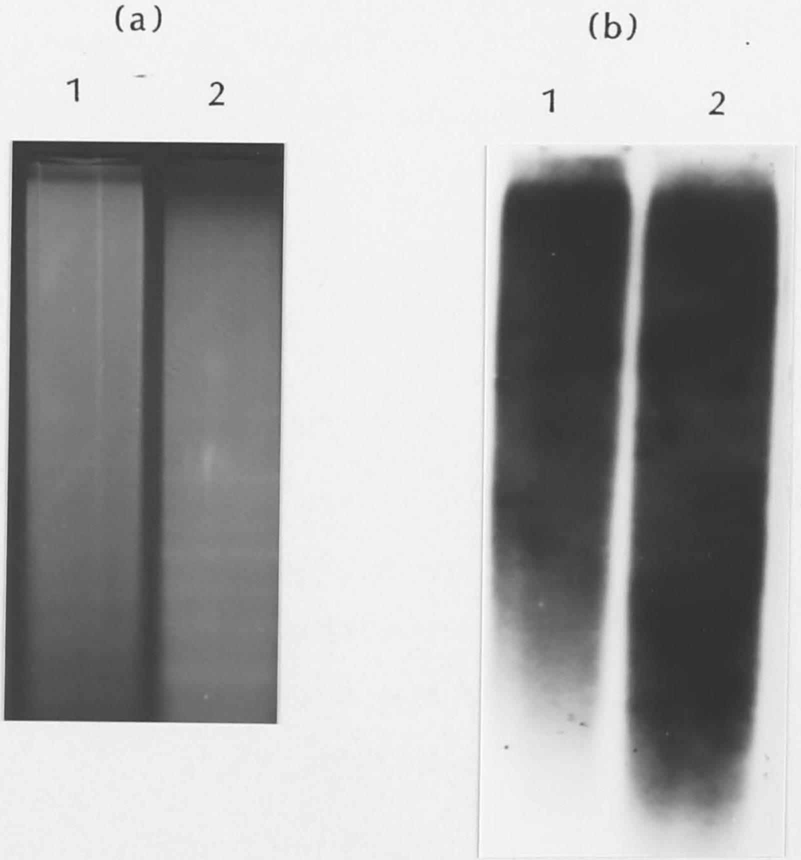


FIGURE 3.6 The location of repeated DNA sequences within the *topaz* region.

(a) Southern blot analysis

Left: EtBr stained gel. DNA from λto6 digested with the following restriction enzymes (left to right): S, SalI; S/B, SalI+BamHI; B, BamHI; S/E, SalI+EcoRI; E, EcoRI; E/B, EcoRI+BamHI; S/H, SalI+HindIII; H, HindIII; H/B, HindIII+BamHI; H/E, HindIII+EcoRI.

Marker - λ/HindIII.

Right: Autoradiogram of a Southern blot of this gel probed with total genomic DNA from SWT.

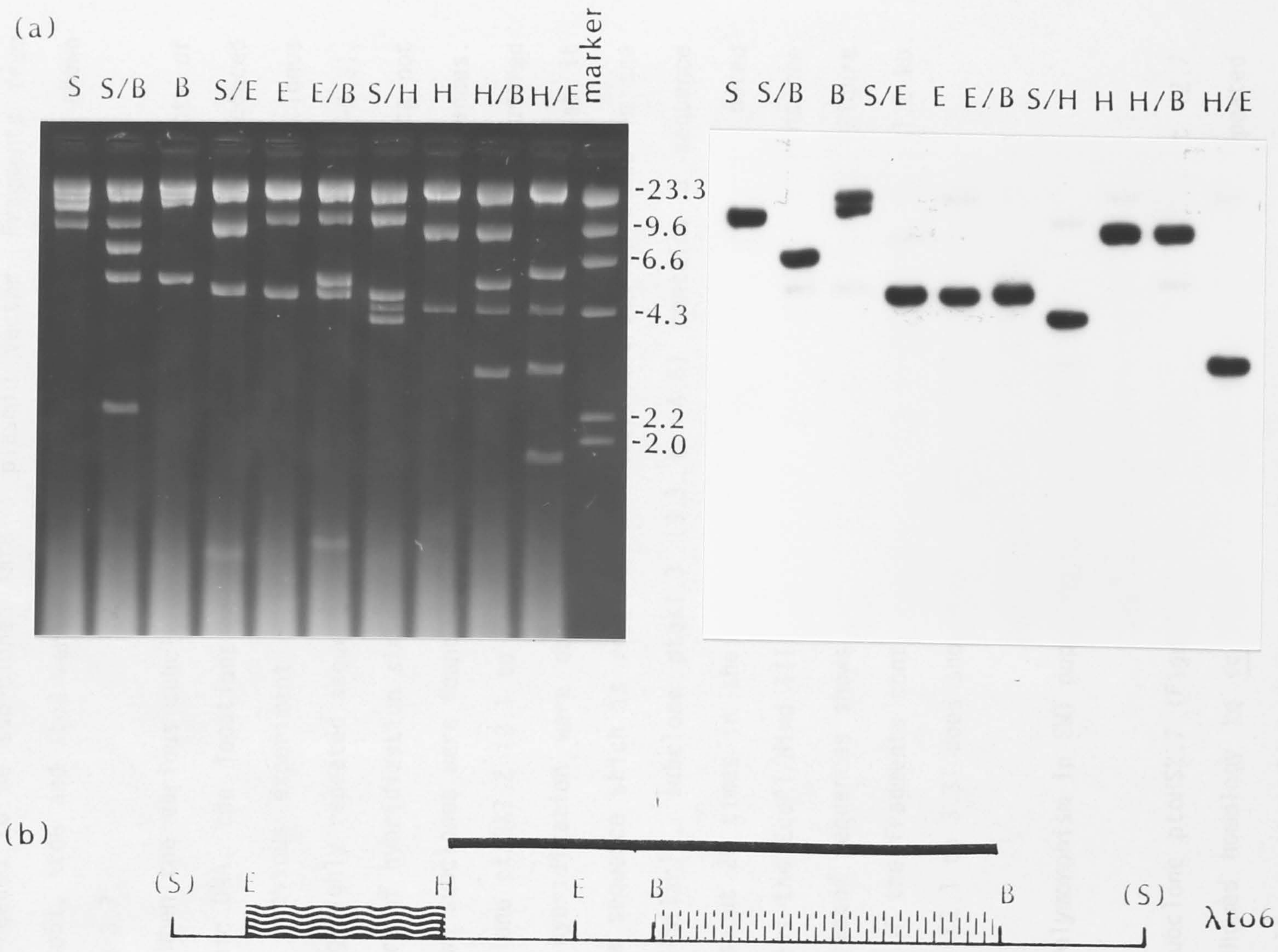
(Nick-translated probe; high stringency hybridization).

(b) Positions of repeats within the λto6 clone.

≋ position of highly repeated sequences

┆ position of moderately repeated sequences (putative)

The solid line above the map indicates region of homology to *scarlet* sequences.


(a)
S S/B B S/E E E/B S/H H H/B H/E marker
-23.3
-9.6
-6.6
-4.3
-2.2
-2.0
S S/B B S/E E E/B S/H H H/B H/E
(b)
(S) E H E B B (S)
λto6

the repeated sequences themselves could be done, the DNA of the topaz λ-clones needed to be sub-cloned into a plasmid vector. Fragments from phages λto51, λto6 and λto8 were subcloned into the vector pUC13 as shown in Figure 3.7.

When the various subclones were used to probe Southern blots of SWT genomic DNA, the locations of the repeated sequence DNA (as deduced from the previous experiment - Figure 3.6b) were confirmed. Subclones containing highly repeated sequences, for example ptoEH3.2 (-4.0 to -0.8), gave smears of hybridization similar to that shown in Figure 3.5 (data not shown). Two subclones were found to contain moderately repeated sequences. When subclone ptoBX1.2 (2.1 to 3.3) was used as the probe, a few hundred bands of hybridization were obtained (Figure 3.8a), indicating that it contains a sequence which is repeated at least hundreds of times in the genome. Similarly, subclone ptoX1.3 (3.3 to 4.6) contains a sequence repeated about 20 times in the genome (Figure 3.8b). It should be noted that whereas the EcoR1/Hind III (-4.0 to -0.8) fragment that contains the highly repeated sequences shows no homology to scarlet sequences (Figure 3.4), one of the fragments containing moderately repeated DNA, the 1.2 kb BamH1/Xba1 (2.1 to 3.3) does show homology to a region of scarlet.

3.2.4 Polymorphism in SWT populations

Subclone ptoHS2.7 (Figure 3.7) carries the largest fragment (2.7 kb) that showed homology to scarlet DNA but which contained no repeated sequences. It was therefore chosen for studies on the genomic organization of the topaz sequences. It was used to probe a genomic blot of SWT DNA digested with the restriction enzyme HindIII (Figure 3.9). From the restriction maps of the topaz clones it was predicted that an 11 kb HindIII

FIGURE 3.7 The subcloning of fragments from the topaz λ clones.

The restriction maps of clones λto51, λto6 and λto8. Underlined are the fragments from each that have been subcloned into the vector pUC13. The subclones have been designated according to the restriction sites by which they were generated and their size in kb.

Restriction enzymes used S, SalI; E, EcoRI; B, BamHI; H, HindIII; X,XbaI.

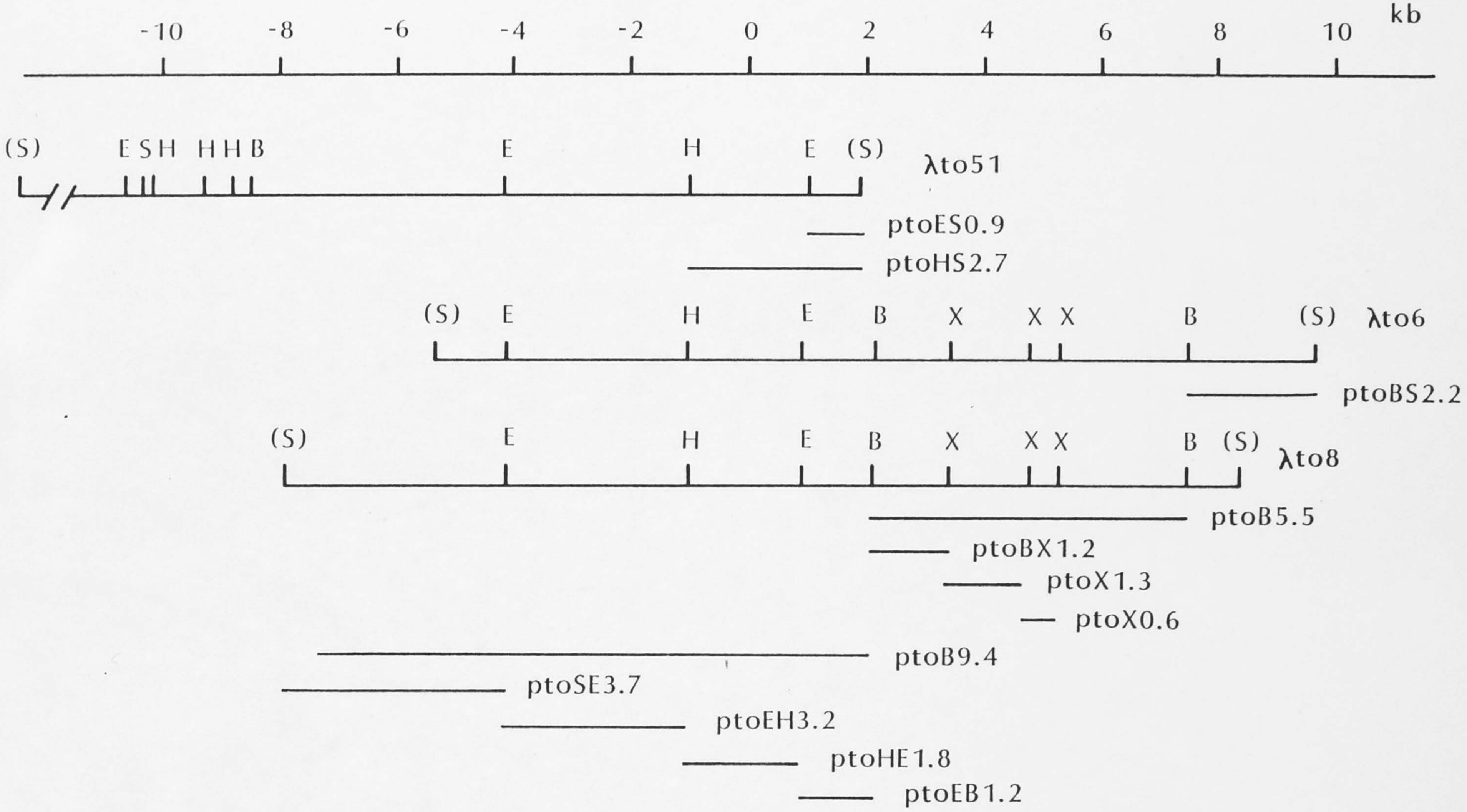

-10
-8
-6
-4
-2
0
2
4
6
8
10
kb
(S) E S H H H B E H E (S)
λto51
ptoES0.9
ptoHS2.7
(S) E H E B X X X B (S)
λto6
ptoBS2.2
(S) E H E B X X X B (S)
λto8
ptoB5.5
ptoBX1.2
ptoX1.3
ptoX0.6
ptoB9.4
ptoSE3.7
ptoEH3.2
ptoHE1.8
ptoEB1.2

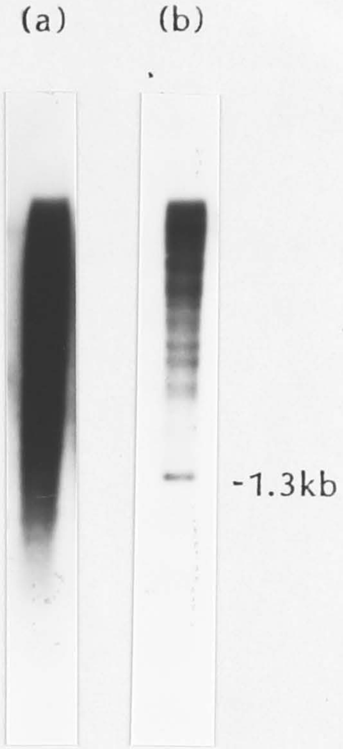
FIGURE 3.8 Repeated sequences within the *topaz* clones. Hybridization of DNA from subclones of the *topaz* region to Southern blots of genomic DNA from SWT.

(a) Autoradiogram of a Southern blot of SWT DNA digested with the restriction enzymes EcoR1 and XbaI and probed with subclone ptoB/X1.2.

(b) Autoradiogram of a Southern blot of SWT DNA digested with the restriction enzyme XbaI and probed with subclone ptoX1.3.

(M13 probes; high stringency hybridization).

Bottom: Restriction map of clone λto8. Positions of subclones used are indicated by dotted lines. Regions of homology with *scarlet* sequences are indicated by solid line.


(a)
(b)
-1.3kb

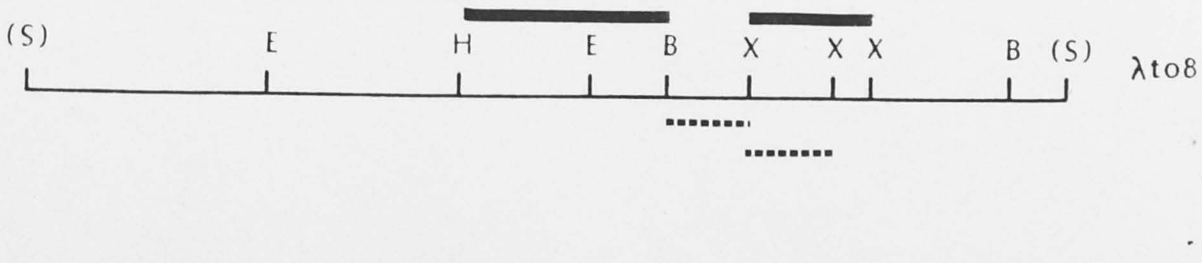

(S)
E
H
E
B
X
X
X
B
(S)
λto8

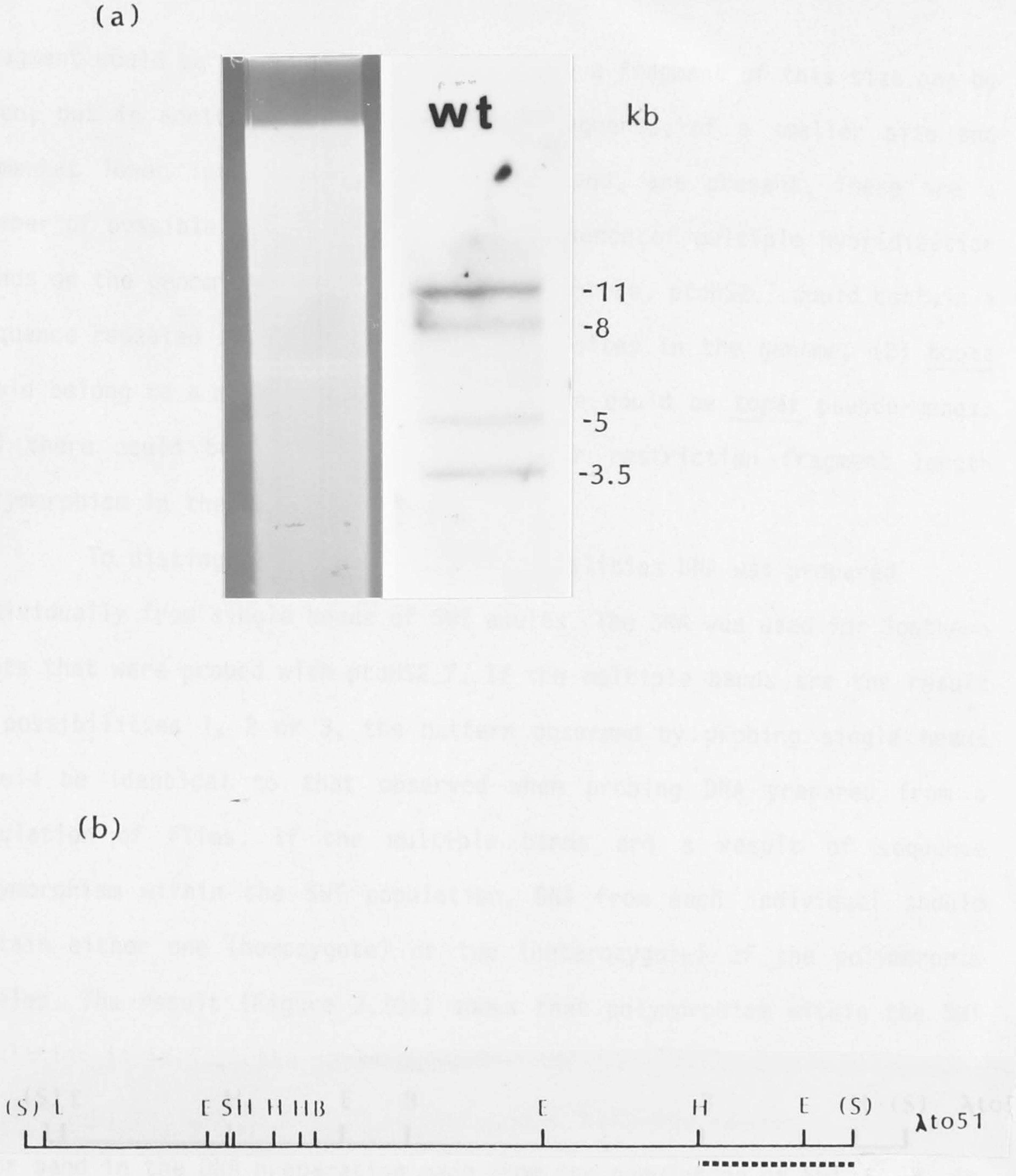
FIGURE 3.9 Genomic organization of unique *topaz* sequences. Hybridization of subclone ptoHS2.7 to a Southern blot of genomic DNA from SWT.

(a) Left: EtBr stained gel of SWT DNA digested with the restriction enzyme HindIII.

Right: Autoradiogram of a Southern blot of this gel probed with subclone ptoHS2.7.

(Nick-translated probe; high stringency hybridization).

(b) Restriction map of clone λto51. The position of the sequences used as a probe in (a) are indicated by the dotted line.


(a)
wt
kb
-11
-8
-5
-3.5
(b)
(S)
L
E
S
H
H
H
B
E
H
E
(S)
λto51

fragment would be produced. Hybridization to a fragment of this size can be seen, but in addition other hybridizing fragments, of a smaller size and somewhat lower intensity than the 11 kb band, are present. There are a number of possible explanations for the presence of multiple hybridization bands on the genomic Southern blot: (1) the probe, ptoHS2.7 could contain a sequence repeated a few times at different sites in the genome; (2) _topaz_ could belong to a multigene family; (3) there could be _topaz_ pseudo-genes; (4) there could be restriction site and/or restriction fragment length polymorphism in the SWT DNA.

To distinguish between these possibilities DNA was prepared individually from single heads of SWT adults. The DNA was used for Southern blots that were probed with ptoHS2.7. If the multiple bands are the result of possibilities 1, 2 or 3, the pattern observed by probing single heads should be identical to that observed when probing DNA prepared from a population of flies. If the multiple bands are a result of sequence polymorphism within the SWT population, DNA from each individual should contain either one (homozygote) or two (heterozygote) of the polymorphic alleles. The result (Figure 3.10a) shows that polymorphism within the SWT population is in fact the explanation for the multiple bands. Each head DNA has a different hybridization pattern, some that are not even seen as a major band in the DNA preparation made from the population of SWT flies. In this experiment, the flies used for the single head DNA preparations and the flies used for the large scale DNA preparation were derived from the same batch (collected at one time during 1983 from one cage and stored frozen).

A similar experiment was repeated about 1 year later, using SWT DNA that was collected on two different occasions. One collection (1984) was used to prepare DNA on a large scale, while the other (1985) was used

FIGURE 3.10 DNA polymorphisms in SWT populations. Hybridization of sequences from the topaz region to Southern blots of genomic DNA from SWT.

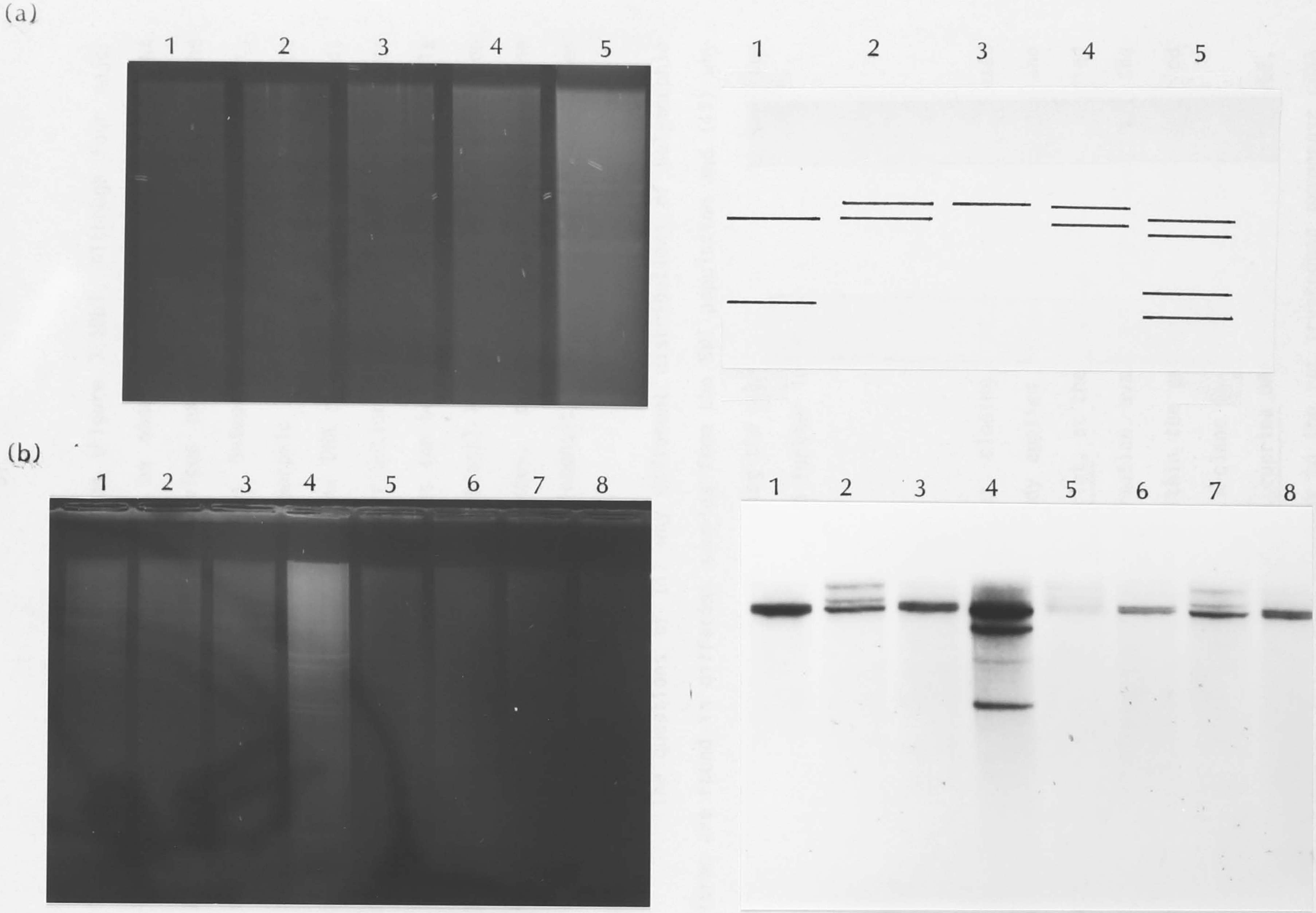
(a) Left: EtBr stained gel of SWT DNA prepared from single heads (lanes 1,2,3,4) or prepared on a large scale (lane 5) digested with the restriction enzyme HindIII.

Right : A diagram of an autoradiogram of a southern blot of this gel probed with subclone ptoHS2.7

(b) EtBr stained gel of SWT DNA prepared from single heads (lanes 1,2,3,5,6,7,8) or prepared on a large scale (lane 4), digested with the restriction enzyme HindIII.

Right: Autoradiogram of a Southern blot of this gel probed with subclone ptoHE1.8.

(M13 probe; high stringency hybridization).


(a)
1
2
3
4
5
1
2
3
4
5
(b)
1
2
3
4
5
6
7
8
1
2
3
4
5
6
7
8

for the single head DNA preparation (Figure 3.10b). Although four major hybridizing polymorphic forms can be observed in the large scale DNA preparations (as seen in the previous experiment - Figure 3.10a), the distribution of signal intensities between these forms is different, suggesting that the different polymorphic forms are present in different frequencies in the SWT samples. The DNA prepared from the single heads shows yet another, different, distribution of polymorphic forms (Figure 3.10b). Only two polymorphic alleles can be seen in the seven individuals sampled, one of which has an additional HindIII site within the region homologous to the probe sequences, producing two bands on the autoradiogram, each with half the intensity of hybridization of the other polymorphic form.

The questions of (i) why different distributions of polymorphic forms are found in different samples from the SWT population and (ii) why most of the isolated λ-clones carrying the topaz region belong to only one polymorphic form, will be considered further in the Discussion.

3.2.5 Yellowish

Preliminary work aimed at cloning the yellowish gene used essentially the same approach already applied in the cloning of topaz and white. Again, as in the case of scarlet, at the time this work was carried out, there was only limited information available with regard to the boundaries of the vermilion gene within the genomic clone it was carried in; information about the vermilion λ-clone is presented in Figure 3.11a.

Southern blots of genomic L.cuprina DNA probed with pvE1 or pvE2, the two fragments which show homology to the rat tryptophan oxygenase cDNA sequences (see Section 3.1), showed no hybridization signal. In a similar

FIGURE 3.11 Restriction maps of the *vermilion* and putative *yellowish* clones and their regions of shared homology.

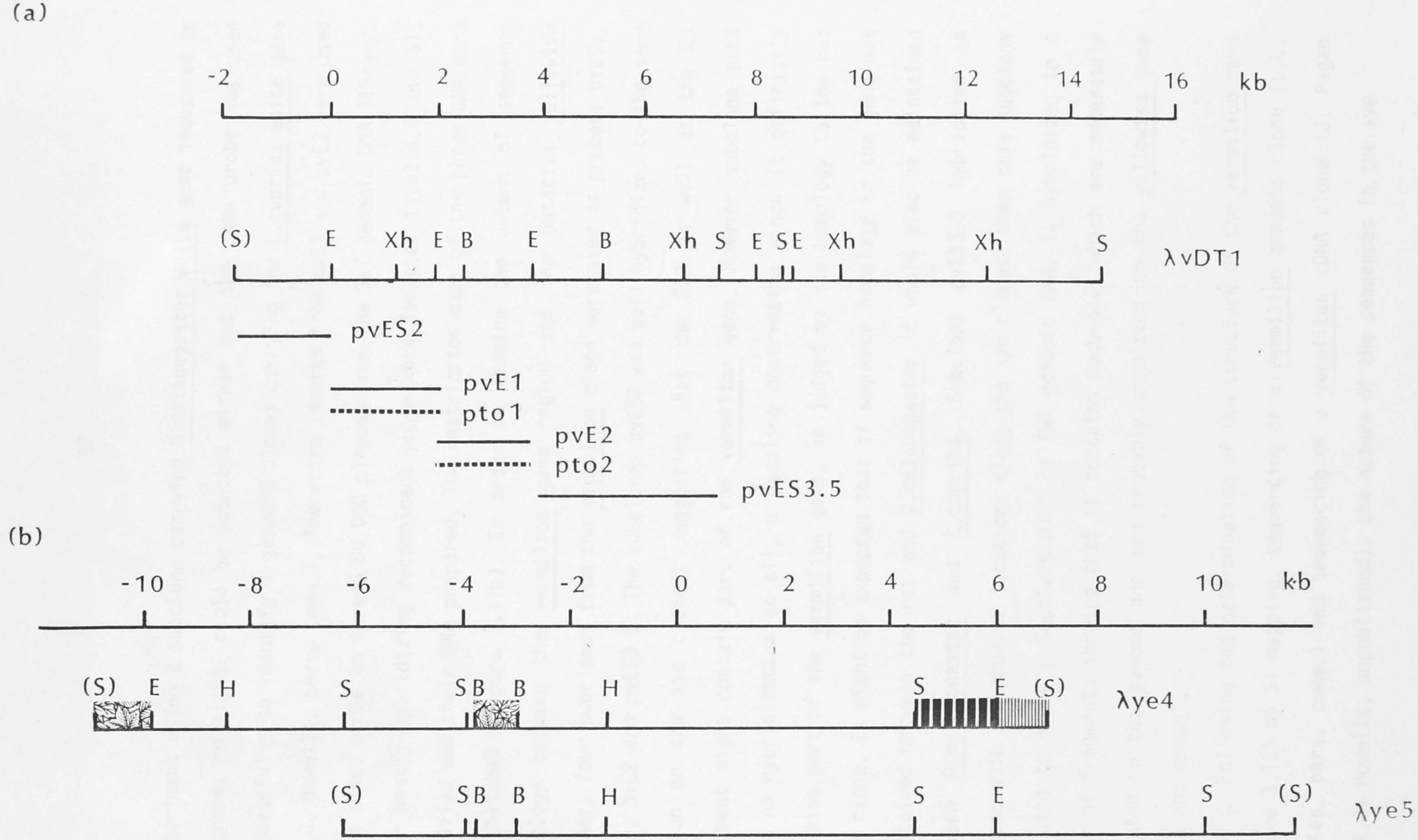
(a) The *vermilion* λ-clone, λvDT1 (from Walker *et al*., 1986a). Regions of homology to the rat tryptophan oxygenase cDNA clones (pto1 and pto2), and sub-clones derived from the λ-clone are marked.

(b) The putative *yellowish* clones. Regions of homology to fragments from the *vermilion* clone are marked

▮▮▮▮▮	homology to pVES2
▥	homology to pVE1
▩	homology to pVES3.5

(clones λye2 and λye6 are contained within the sequences carried in clones λye4 and λye5 and therefore their restriction maps are not shown separately).

Restriction enzymes used: S, SalI; E, EcoRI; B,BamHI; H, HindIII; Xh, XhoI.


(a)
-2
0
2
4
6
8
10
12
14
16
kb
(S)
E
Xh
E
B
E
B
Xh
S
E
SE
Xh
Xh
S
λvDT1
pvES2
pvE1
pto1
pvE2
pto2
pvES3.5
(b)
-10
-8
-6
-4
-2
0
2
4
6
8
10
kb
(S)
E
H
S
SB
B
H
S
E
(S)
λye4
(S)
SB
B
H
S
E
S
(S)
λye5

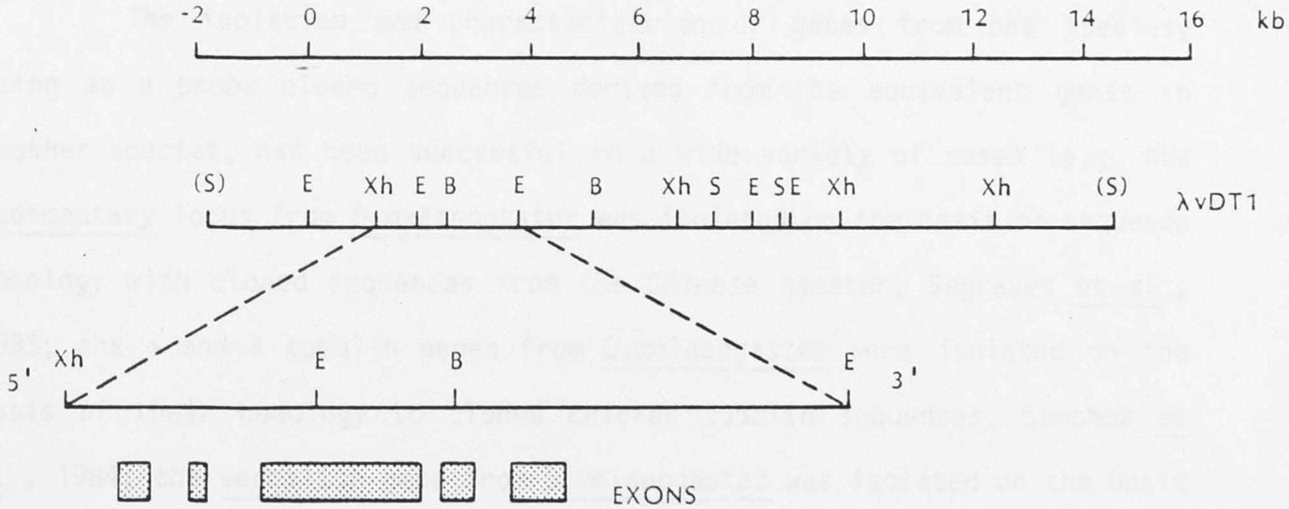
experiment using a subclone carrying *D.melanogaster* *white* gene sequences as a probe, no signal could be detected either but the same probe was used successfully to identify λ genomic clones carrying the *L.cuprina* *white* gene (A.J. Howells, pers. comm.). Therefore inserts from pvE1 and pvE2 were used as a mixed probe to screen 80,000 plaques from the SWT genomic DNA library. Four weakly hybridizing recombinant phage were isolated (λye2,4,5 and 6), purified and their DNA prepared. The restriction maps of the phage DNA were established (Figure 3.11b). In order to examine the extent of sequence homology between the *vermilion* gene region and the putative *yellowish* clones, four subclones from the *vermilion* clone were used as probes; pvES2, pvE1, pvE2 and pvES3.5. The subclones pvES2 and pvE1 hybridized to the same region on the λye clones, suggesting that the ES2 as well as the E1 fragment might contain part of the *vermilion* gene. However subclone pvE2 gave no hybridization at all, a puzzling observation since it definitely contains part of the *vermilion* gene, as judged by its homology to the rat cDNA clone. It might be expected that if sequence homology in the gene was maintained between the rat and *D.melanogaster* it would also be maintained between *D.melanogaster* and *L.cuprina*. Subclone pvES3.5 hybridized to sequences in a number of locations along the λye clones. When this subclone was used to probe a Southern blot of SWT genomic DNA, it hybridized to a smear of fragments showing that it contains sequences which are moderately repeated in the genome and are probably unrelated to the *yellowish* gene (data not shown).

Following the determination of the structure of the *vermilion* gene (Figure 3.12) by S1 mapping, sequencing of a *vermilion* genomic clone (L.L. Searles, pers. comm.) and sequencing of a *vermilion* cDNA clone (J. Pagan and A.J. Howells, unpublished), the nature of the sequences in the λye

FIGURE 3.12 The structure of the vermilion gene.

The positions of the exons and introns in the vermilion gene (L.L. Searles; J. Pagan pers. comm.). The exons are marked by the solid blocks and the introns by the gaps between them.

Restriction enzymes used: E, EcoRl; B, BamHl; Xh, XhoI.



clones was re-examined in our laboratory. Dr Pagan* used the vermilion cDNA clone to probe Southern blots of DNA from the λye clones, SWT genomic DNA and the SWT genomic library. The results showed (i) that the λye clones do not share any sequence homology with the vermilion cDNA clone; (ii) that a clear band of hybridization can be detected in the Southern blots of the L.cuprina genomic DNA with the cDNA probe and (iii) that no positively hybridizing phage were found in the genomic library screen (80,000 plaques were probed). The interpretation of these results will be presented in the following Discussion.

3.3 DISCUSSION

The isolation and characterization of genes from one species, using as a probe cloned sequences derived from the equivalent genes in another species, has been successful in a wide variety of cases (e.g. the rudimentary locus from D.melanogaster was isolated on the basis of sequence homology with cloned sequences from the Chinese hamster, Segraves et al., 1983; the α and β tubulin genes from D.melanogaster were isolated on the basis of their homology to cloned chicken tubulin sequences, Sanchez et al., 1984; the vermilion gene from D.melanogaster was isolated on the basis of its homology to cloned rat tryptophan oxygenase sequences, Walker et al., 1986; the human rhodopsin gene was isolated on the basis of its homology to cloned bovine rhodopsin sequences, Nathans et al., 1986). The

* The experiments performed by Dr Pagan (involving the vermilion cDNA and the λye clones) were done during a short period in which I introduced her to techniques of molecular biology.

success of this approach probably depends on the evolutionary distance between the two species as well as on the nature of the gene in question.

In the case of *L.cuprina*, the *white* eye colour gene has been isolated on the basis of its homology to the cloned *white* gene from *D.melanogaster* (A.J. Howells, pers. comm.), and in this Chapter the isolation of the *topaz* gene on the basis of its homology to the corresponding *scarlet* eye colour gene has been described. However it is clear that for the isolation of the *yellowish* gene this strategy has failed so far.

3.3.1 *Topaz*

The evidence that indeed the *topaz* gene has been isolated is based on the observed co-linearity of sequence homology between fragments from *scarlet* and *topaz* (Figure 3.4). The probability of obtaining such a pattern of homology between unrelated genes or sequences by chance must be very low.

By using smaller fragments from the *scarlet* and *topaz* genes as hybridization probes, a fine scale analysis of the exact regions of shared homology could be determined. There are however a number of questions concerning the interpretation of the observed homologies which hybridization studies of this sort could not resolve. (i) Do the regions of homology shared between the two genes in effect define the gene boundaries or do they only correspond to parts of the genes which code for areas of a highly conserved function in the protein (such as substrate binding sites or regions which interact with other proteins or membranes), while the rest of the coding region has diverged considerably? (ii) Do the regions of shared homology represent two (or more) genes adjacent in *D.melanogaster* and L.cuprina? This interpretation is raised by the gap in sequence homology along the topaz clone (the 2.1 to 3.3 region), and the presence of repeated sequences in that very same region (Figures 3.6 and 3.8). To answer these questions, as well as to define the precise exon/intron structure of the two genes, detailed nucleotide sequencing of the topaz and scarlet genes was undertaken and will be described in the next Chapter.

3.3.2 Repeated DNA sequences

Evidence for the presence of repeated DNA sequences within a number of fragments along the topaz region has been obtained (Figure 3.6). Work done on two other L.cuprina genes, white (A.J. Howells, pers. comm.) and the cuticle protein genes (Skelly, 1985) has also demonstrated the existence of dispersed repeated sequences, suggesting that they are a general feature of genomic organization in this species. In the case of topaz, the largest DNA fragment shown to be composed entirely of unique DNA is only 2.9 kb long (-0.8 to +2.1). It therefore appears that the L.cuprina genome is organized in a fashion typical of higher eukaryotes where short segments of unique DNA are interspersed with repeated sequence elements, in what is commonly referred to as short periodicity interspersion (Davidson et al., 1973; Jelinek and Schmid, 1982). This forms an intriguing contrast to the long periodicity interspersion pattern characteristic of D.melanogaster (Crain et al., 1976a,b; and see Introduction) particularly in light of the similarities between the two species, such as the conservation of genetic linkage groups and the biochemistry of ommochrome synthesis.

The work described in this Chapter clearly demonstrates the presence of repeated DNA sequences in the topaz gene region and suggests they are relatively short. Detailed information about their structure and relationship to topaz coding sequences can only be obtained by DNA sequencing, as will be described in Chapters 4 and 6.

3.3.3 Polymorphism

It is clear from the Southern blot analysis of L.cuprina genomic DNA that considerable restriction site and/or restriction fragment length polymorphism exists in the laboratory wild type lines of L.cuprina. DNA polymorphisms in wild type populations have been described in other species (e.g. Orkin et al., 1982; Johns et al., 1983; Kreitman, 1983; Estelle and Hodgetts, 1984), however the high level of polymorphism present within the SWT lines and, more so, the apparent difference in frequencies of the various topaz alleles found in DNA preparations made from different batches of flies requires further comment.

The SWT (standard wild type) line was generated by Dr G. Foster about fifteen years ago with the intention of establishing isogenic lines. Since under laboratory conditions L.cuprina has a 3 week life cycle, 3 population cages, each with about 1000 flies from an LBB wild type line (which was collected in the field around 1950 by Dr L. Barton-Brown) were set up. Each week the flies from one cage would reach maturity and about 30 laying females would be kept to start the next generation. Although the SWT lines are regarded as one line, however, because of the way they are maintained they have almost become three independent lines - almost, but not strictly so, since in practice if one line died or weakened it would be supplemented by flies from the other two lines, but no records were kept of such events. An attempt to isolate isogenic lines by brother-sister pair matings was abandoned because fertility declined rapidly, despite selection in each generation of mating resulting in progeny that exhibited higher than 80% egg to adult survival. This result indicates high levels of recessive lethals in the population (Foster and Whitten, unpublished).

When the above history of the SWT lines became apparent (it had been assumed early in the project that SWT was a single line), it seemed plausible that perhaps each of the DNA preparations which showed variation in frequencies of the different topaz alleles had been made from flies originating from a different SWT line. To study this possibility, flies were collected from the three lines over a three week period (during 1986), their DNA prepared and used in a Southern blot analysis (Figure 3.13). The three 1986 SWT DNA samples appear identical and show virtually no polymorphism. They have only one of the major alleles seen in the 1983 SWT sample (which is the same DNA used for the preparation of the SWT genomic library). The result, which shows there is reduced polymorphism at the topaz locus in the current SWT populations, was quite surprising (particularly in light of some of the sequencing data that will be presented in Chapter 6) and suggests that at least one of the SWT lines might have gone through a "genetic bottle-neck" (and been replaced by flies from the other line/s) during the past three years. It would be interesting to compare DNA preparations from the LBB lines (which are also kept in three population cages) to see if differences in topaz alleles frequency can be detected there.

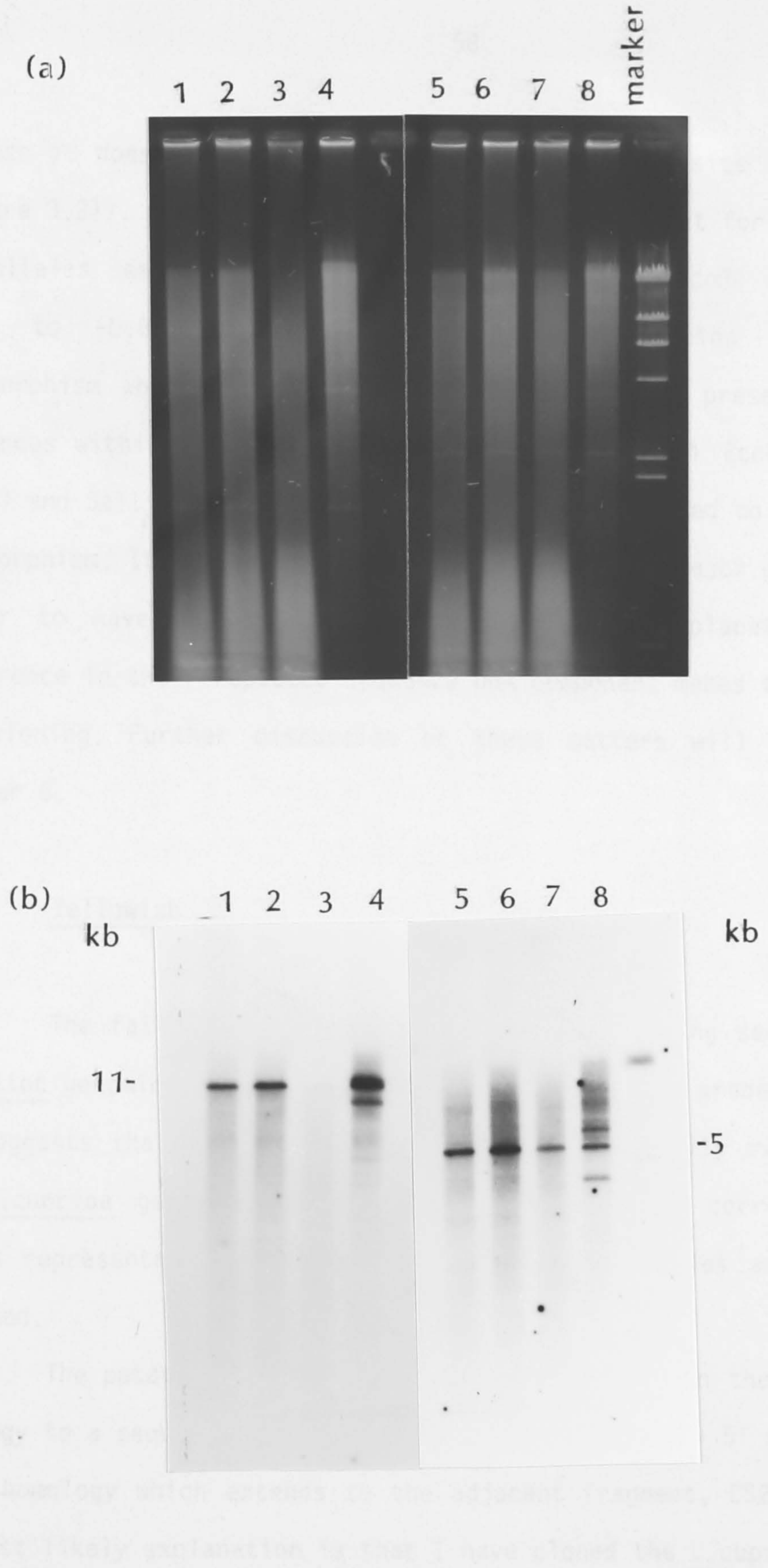
It has been noticed, that although both the LBB and SWT DNA used in preparation of the L.cuprina genomic libraries were polymorphic at the topaz gene region (see Figure 3.13 for SWT, data not shown for LBB), little polymorphism within the putative coding region could be detected in the topaz λ-clones. Clone λto52 can account for the 11 kb Hind III allele (Figure 3.9) but apart from λto51 (for which our knowledge is incomplete

FIGURE 3.13 Changes in the level of DNA Polymorphism in SWT populations.

(a) EtBr stained gel of SWT DNA, prepared from flies originating from the three separate lines in 1986 and from one line in 1983 (lanes 4 and 8), digested with the restriction enzyme HindIII (lanes 1-4) and EcoRI (lanes 5-8). Marker - λ/HindIII.

(b) Autoradiogram of a Southern blot of gel (a); lanes 1-4 were probed with subclone ptoES0.9 (left) and lanes 4-8 with subclone ptoHE1.8 (right).

(M13 probes; high stringency hybridization).


(a)
1 2 3 4 5 6 7 8 marker
(b)
1 2 3 4 5 6 7 8
kb
kb
11-
-5

because it does not extend much beyond the HindIII site at position -0.8 (Figure 3.3)), none of the other λto phage can account for the smaller Hind III alleles seen on the blot. The shorter HindIII/EcoRl fragment in λtol (-3.6 to -0.8) can be attributed to restriction fragment length polymorphism which is perhaps associated with the presence of repeated sequences within this fragment, while the additional EcoRl (-3.0), BamHl (-1.9) and SalI (+4.9) sites in λtol2 can be attributed to restriction site polymorphism. It is not clear why none of the other major polymorphic forms appear to have been cloned. The most likely explanation is that a difference in their repeated sequence DNA component makes them unfavourable for cloning. Further discussion on these matters will be presented in Chapter 6.

3.3.4 *Yellowish*

The failure to clone the *yellowish* gene using sequences from the *vermilion* genomic clone or from the *vermilion* cDNA as probes was unexpected and suggests that our SWT genomic library is not fully representative of the *L.cuprina* genome, an observation which can be correlated with the biased representation of the *topaz* polymorphic alleles among the clones isolated.

The putative *yellowish* clones were selected on the basis of their homology to a sequence which appears to lie immediately 5' to the *vermilion* gene, homology which extends to the adjacent fragment, ES2 (Figure 3.11): The most likely explanation is that I have cloned the *L.cuprina* counterpart of the gene (of unknown function) which lies next to *vermilion*. Despite close proximity of these sequences to the *vermilion* gene in *D.melanogaster* (within the 1.3 kb of EcoRl/XhoI fragment), in *L.cuprina* it hybridizes to unique sequences (since only one set of overlapping clones was isolated) which are not immediately adjacent to yellowish. Based on the restriction maps of the λye clones and on the regions of homology between the L.cuprina and D.melanogaster sequences (Figure 3.11), there are at least 5kb of other DNA sequences between this unidentified gene (or sequence) and the yellowish gene (provided the Southern blot result showing homology between the vermilion cDNA clone and L.cuprina genomic sequences is due to actual homology with yellowish sequences and is not fortuitious). This indicates that the distance between genes in L.cuprina might be much longer than in D.melanogaster, an observation which, if generally true, could partially account for the difference in genome size between the two species. It should be pointed out that the sequence in question is not necessarily even on the same chromosome as the yellowish gene and the speculations above are only relevant if the rather generalised observations on the conservation of genetic linkage groups between the two species (see Introduction) extends to the fine scale of individual adjacent genes.

The DNA fragment carried in subclone pvES3.5 which is located 3' to the vermilion gene, is present as a unique sequence in D.melanogaster. In L.cuprina however it hybridizes to a sequence which is moderately repeated in the genome, one copy of which is adjacent to the sequence homologous to pvES2 (Figure 3.11). In D.virilis, pES3.5 also hybridizes to a moderately repeated sequence (M.A. Rye and A.J. Howells, unpublished). Whether it is actually the same sequence from within the 3.5 kb fragment which is repeated in L.cuprina and D.virilis is unknown but it would be interesting to find this out as a possible clue towards understanding the significance of differences in the short dispersed repeat components of the genome between these species. D.melanogaster, D.virilis and L.cuprina could be suitable species on which to make such comparisons since the evolutionary distance thought to exist between *D.melanogaster* and *D.virilis* (approximately 60 million years) is about half that between *D.melanogaster* and *L.cuprina* (approximately 100 million years) (Beverley and Wilson, 1984).

3.3.5 Evolution of eye colour genes

Out of the three homologous pairs of eye colour genes examined, *scarlet* and *topaz* show the highest level of sequence conservation as judged by the appearance of hybridization signals when using *scarlet* sequences to probe Southern blots of *L.cuprina* genomic DNA. *Vermilion* and *yellowish* show less sequence homology as increased probe sensitivity, utilizing *vermilion* cDNA sequences, was required before *yellowish* sequences could be detected on Southern blots. This result also emphasises the increased resolution achieved by using M13 probes compared with nick translated probes. So far sequences from the *white* gene of *D.melanogaster* have not detected *L.cuprina* *white* gene sequences by Southern blot hybridization; however the use of M13 probes has not yet been attempted.

Although the studies reported are still preliminary, they appear to provide evidence that three *L.cuprina* genes associated with the same biochemical pathway have diverged to different extents from their homologous genes in *D.melanogaster*. Further studies will be required to establish the pattern of divergence, to examine whether it affects certain parts of the gene more than others or whether it reflects a greater extent of silent base substitutions in one gene compared with another. Some of these questions will be addressed in the next Chapter.

CHAPTER 4

THE ANALYSIS OF THE TOPAZ GENE REGION BY DNA SEQUENCING

4.1 INTRODUCTION

The determination of the exon/intron structure of a gene is a pre-requisite before questions concerning its regulation, evolution and structure in relation to that of the genome can be addressed. The most unequivocal way to establish gene structure is to compare the sequence of the RNA encoded by the gene (usually via a cDNA) or the amino acid sequence of its protein to the nucleotide sequence of the gene itself and so identify the coding (exons) and non-coding (introns) regions. When the protein sequence is unknown or a full length cDNA clone is unavailable, other approaches can be used.

The establishment of the structures of the white and vermilion genes from D.melanogaster are perhaps the two examples most relevant to the work described in this thesis. A genomic clone carrying the white gene was sequenced by O'Hare et al. (1984); the 3' end of the gene was defined by comparing the genomic sequence with the sequence of a truncated cDNA clone derived from that region. The rest of the coding regions were established by hybridizing short subclones from along the gene to polyA$^+$ RNA. Since such hybridization studies would not detect the presence of short introns, the DNA sequences of those subclones which hybridized to the RNA were searched for probable protein coding regions and potential exon/intron splice sites. The putative structure presented by O'Hare et al. (1984) was proposed on the basis of the combination of the above data. Subsequently, on the basis of sequencing a much longer cDNA clone and other white genomic clones, this structure was found to have only one error which resulted from an incorrect assignment of one splice junction (S. Mount, pers. comm.).

The vermilion gene structure (as discussed in the previous Chapter) was determined by L.L. Searles (pers. comm.) by sequencing a genomic clone carrying the gene followed by S1 nuclease mapping. The proposed structure was confirmed by the subsequent sequencing of a vermilion cDNA clone (J. Pagan and A.J. Howells, unpublished); however this cDNA clone did not contain the complete transcript so the structure of the 5' region is still uncertain.

The structure of the topaz gene was determined by a novel approach, as will be described in this Chapter. Essentially, this involved simultaneously sequencing the homologous gene region from two distantly related species (topaz from L.cuprina and scarlet from D.melanogaster) and comparing the DNA sequences for conserved regions.

4.2 RESULTS

The topaz gene region was subcloned into phage M13 vectors and sequenced by the dideoxy chain termination procedure (see Chapter 2). The sequencing strategy is shown in Figure 4.1. The scarlet gene was sequenced in conjunction with topaz largely by Mrs D. Boyle (described in Tearle, 1986). The computer analysis of the sequences from topaz and scarlet was done in collaboration with R.G. Tearle.

4.2.1. Gene structure

The putative exon/intron structure of the topaz and scarlet genes was established by considering a number of criteria; essentially, regions

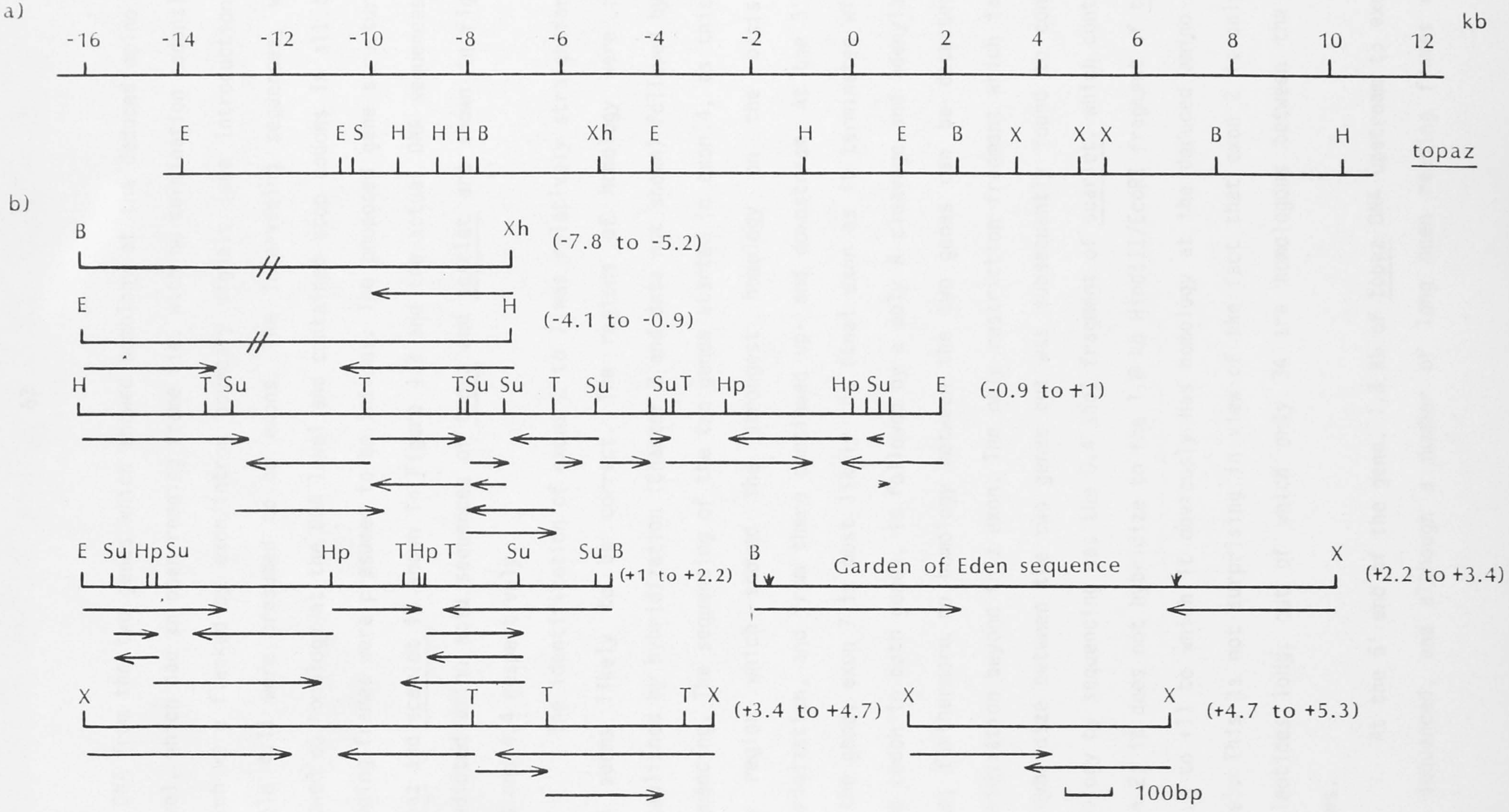
FIGURE 4.1 Sequencing strategy for the _topaz_ gene region

(a) Restriction map of the _topaz_ gene region.

(b) Sequenced regions of subclones from _topaz_. Arrows indicate starting points and direction in which sequences were read from the M13 recombinants. Length of arrows indicate the extents to which sequences were read from single reactions.

Restriction enzymes used: E, EcoR1; B, BamH1; H, HindIII; S, SalI; X, XbaI; Xh, XhoI; Hp, HpaII; Su, Sau3A; T, TaqI.

Examples of sequencing gels are presented in Chapter 6.


a)
kb
-16
-14
-12
-10
-8
-6
-4
-2
0
2
4
6
8
10
12
E
E S
H
H H B
Xh
E
H
E
B
X
X X
B
H
topaz
b)
B
Xh
(-7.8 to -5.2)
E
H
(-4.1 to -0.9)
H
T Su
T
TSu Su
T
Su
SuT
Hp
Hp Su
E
(-0.9 to +1)
E Su HpSu
Hp
THpT
Su
Su B
(+1 to +2.2)
B
Garden of Eden sequence
X
(+2.2 to +3.4)
X
T
T
T X
(+3.4 to +4.7)
X
X
(+4.7 to +5.3)
100bp

of DNA from the two genes which showed homology at the deduced amino acid level, which had an open reading frame (ie. with no termination codons) and which were flanked by exon/intron boundary signals (see Introduction and Table 4.1) were presumed to be exons. The intervening sequences, which showed no homology at the DNA level and contained stop codons in all three reading frames were presumed to be introns. The proposed gene structure of topaz and scarlet is shown in Figure 4.2 and the actual DNA sequences and predicted amino acid sequences of topaz and scarlet are shown in Figures 4.3 and 4.4 respectively.

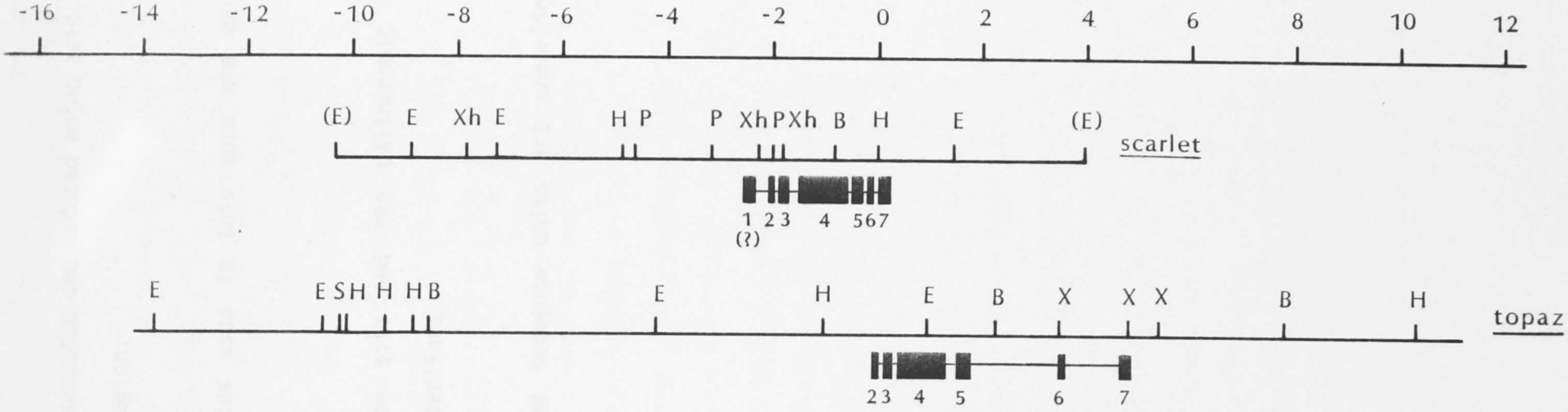
The identification of exons 2 to 7 was relatively straightforward and seems likely to be correct. The regions of homology were first established by hybridization (Chapter 3 and data not shown) followed by DNA sequencing. The sequencing of the two genes started in exon 4, as this was the region which showed the strongest homology on the basis of hybridization, and from there continued up- and downstream. At the 3' end of the gene, exon 7 is most likely the final exon as it terminates with a stop codon in both genes, is followed by a poly A cleavage and adenylation site; furthermore, no homology between the two genes can be detected by hybridization beyond this exon. The only restriction fragment which failed to hybridize between the two genes but was subsequently found to contain homology by sequencing was the 0.4 XhoI fragment of scarlet, which contains exon 2; it does not hybridize to the 1.8 kb HindIII/EcoRI fragment of topaz (-0.8 to +1) to which it obviously has homology at the deduced amino acid level. This is not surprising in view of the fact that exon 2 is only 49 nucleotides long, out of which only 34 are homologous between the two genes.

At the 5' end of the gene, 1.3 kb of topaz DNA upstream to exon 2 was sequenced, and although a number of long open reading frames were

FIGURE 4.2 The putative structure of the topaz and scarlet genes.

The position of the exons is marked by the boxed areas; the introns are marked by lines.

Restriction enzymes used: E, EcoRl; B, BamHl; H, HindIII; S, SalI; X, XbaI; Xh, XhoI; P, PstI.



-16
-14
-12
-10
-8
-6
-4
-2
0
2
4
6
8
10
12
(E)
E
Xh
E
H
P
P
Xh
P
Xh
B
H
E
(E)
scarlet
1
(?)
2
3
4
5
6
7
E
E
S
H
H
H
B
E
H
E
B
X
X
X
B
H
topaz
2
3
4
5
6
7

FIGURE 4.3 The DNA nucleotide sequence and deduced amino acid sequence of the topaz gene region.

The restriction sites used in this work are marked and underlined.

The poly A addition site and the palindromic sequence preceding it are underlined.

Examples of repeated sequence units are underlined (see also Table 4.6).

```
-5400 CCGATTGGTATCTTTCAAGTTTGGTATAGAACCATGTTATTGTGCGTTAATAACAAACCC

      AATATAACCCTCCCACTAAAGTGGTCATCAGGGTAATAAATAGAAATAAATATAAATAAC

      AAAAATAAGTAATAGTTATTCACAAGAATTACTAACACTTGTCATGACAAGTATACAGTA

      ACTGAAATGAGTATGGGATTTTAGAGAACAATCGCGAGCCATTTGGCTCATGATGTTCTT

      CCCTTACTTAAAAAAAATTAATTAATTTTCTTTATTGTTATTGATCCTCGAC
       (Gap of c. 1000 bp)

-4100 TATGGACTAGTCTACTGGTTAGTCTATTGCCTAGTCTATAGTTTAGTTTATTGACTAGTT

      TATTTATAAGTCTATGAACTTATTAATTAATTAGTTTATGCATATTGGCTAGTCTACTGA

      CTAGCTATTGTTTAGTCTATGGACTAGTTTATTGTCTAGTCTATGGACTAGTTTATTGTC

      TAGCTATTTTCACGCTATCCACTCGCTATGGACTGGTAAGCTATTGCCTAGCAAGTTAGG
      (Gap of c. 2600 bp)
                                                    ATTCATATTTTGTT

      GTTTGTTATTTATTTTCTTTTGAGTTTTCTAATGACTTTTATTTAAATGCATAATAAAAG

      AGAAGTTAAATTATTGGTTTTGTGATAAAAAACAAATAAGATGTGTAATGTAATAGTTTT

-1100 CAATTATAGAGACGTAAATTTACAGAGAATTAACCGTGTGTGTGTCATGCAACATTGTAA

      TGAATGATAAATGTTTTTGAGCTGTAAATAAACATTGTAACCCAGTGAACATATTTTAAC

      TAATTTATAAAAAAACAAGTAGAATGTATAGTCGGGCGTGACTGACTATATGATATTTTT

      AAAAATAAAGCATCTATGTTGAAGTTTTCCTTAATTTGGATATATAGTTTTGATGAAAAA
                  -       HindIII
      ATGATTTTCTGATGGCATCTAAGCTTAATTGTGTACTGTTGACATTAAAGTTAAATTTTC

-800  CGAATTTCATTGAAATACAGGTGCTTATGCAGAAAATTAATGAAGAGGAGGAGTTGCAAG

      CGTTAAATAAAGGCCGCTCATTGCGAAATTTTGCAATATCATTTATTGCGATTTAAAACG

      TAGTTGTGCATATTTTTGTTAGGATATTAGTATAATTGACATGATTATGAGCTCTGGGGT

      TAGGGGTCCGTTTGTGTGGGTGCTAGGTGAAGGAATAAATCGATATCGTCCAAATTCAAT
                                           Sau3A
      AGTATTCGTGCTTTTCACTCAATTATGTTTAAAACTGCGATCTGTAGTTTGATGATTCTG

-500  ACTTCAAAAAATGAAACATTTATCAATACTTCTAAAGTAGGACTAAATGAATTTTTTACC

      TTCTGTGGAAGTGATATTTATTGACTTCCAAAGAGGCAAGATGTTACTCTTTTTGTAGTA

      AAATTTCTAACTAATTTTATTTTATTTAAATTTATTTATGGAAGTTATTTTTTTTTCTGG

      GTGTGTCTATTGAAATCATTTCTGGTATTTTCGCAATCTAAATAATTCTTTCAGACATAT
                Taql
      GTATGTAGTTTCGAAAAGAGACCTTTTAAAAATCAGCAAAATCGGTCCATACGACGTTAT

-200  GACCAAAAGTTCAAGACAACAACGTATAGAGTTATATGGTAAATTTTCTTGGGCTTAGCT

      TCCTTCTTTATAAAGCAAAACAGAGAATGAATAAAGCATAACATATGAAGCACTACGTTG
```

```
                                  Taql Sau3A
      TTATCTGATATTTCATCGTTTTAGATTTTCGACGATCATGTTGATCTCTTAAAAACGCAA
                                                            Sau3A
       CTTTACTAAAGAAAAAAAGAAAACATTTTAGTGATAAAAATATTTTATTTGATCTATTTT

       TACAGTGGTTCAGGAAAAACAACATTAATGTCTGTACTTGCATATCGCCAACCAGGTAAT
            aGlySerGlyLysThrThrLeuMetSerValLeuAlaTyrArgGlnProV
                                   Taql
+100   TAAAGTTTAGATTTTATTATTTAAATAATCGACAGACGTAGAAACTCGTGAAGACCAAAG


       ATTAGATAATTAAAAATAAAAAAATATATTTCAATAATTTCCAGTTGGTACCGTAGTTCA
                                                   alGlyThrValValGl

       AGGTGATATTCTTATCAATGGCCGACGTATAGGACCATTTATGCATCGCATAAGTGGTTG
       nGlyAspIleLeuIleAsnGlyArgArgIleGlyProPheMetHisArgIleSerGlyCy
                                 Sau3A
       TGTTTATCAAGATGATTTATTTAATGGATCACTTACCGTGGCAGAACATATGCACTTTAT
       sValTyrGlnAspApsGluPheAsnGlySerLeuTyrValAlaGluHisMetHisPheMe

       GGTAGGAGAACTTCATGTCCTTTTAAATTATTTTTAATAATTTTTACATATTTTAGGCTC
       t                                                       AlaL
                  Sau3A                                     Sau3A
+400   TCTTACGTTTAGATCGCCGCGTTAGCAAGCAGGAACGTAAACTTATAATACAAGATCTTT
       euLeuArgLeuAspArgArgValSerLysGlnGluArgLysLeuIleIleGlnAspLeuP

       TCGAACGTACAGGTCTATTGGGTGCTTCTAATACACGTATTGGTTCGGGAGATGATGAAA
       heGluArgThrGlyLeuLeuGlyAlaSerAsnThrArgIleGlySerGlyAspAspGluL

       AAGTGTTATCGGGTGGTGAACGTAAACGTTTAGCTTTTGCTGTGGAATTGTTAAATAATC
       ysValLeuSerGlyGlyGluArgLysArgLeuAlaPheAlaValGluLeuLeuAsnAsnP

       CGGTGATATTATTTTGTGATGAACCCACCACTGGTTTGGATTCTTATAGTGCTCAGCAGT
       roValIleLeuPheCysAspGluProThrThrGlyLeuAspSerTyrSerAlaGlnGlnL

       TGGTGCAAACCCTTTACGATTTAGCCAAAAAGGGTACCACTATCTTATGCACCATACACC
       euValGlnThrLeuTyrAspLeuAlaLysLysGlyThrThrIleLeuCysThrIleHisG

+700   AACCGTCTTCACAATTATTTGATATGTTTAATAATGTTCTCTTTTTGTCGGAGGGCAGAG
       lnProSerSerGlnLeuPheAspMetPheAsnAsnValLeuPheLeuSerGluGlyArgV

       TGGCCTTTACTGGTTCACCACAAAATGCTTTGGATTTTTTTGCTCAAAATGGTTATAGAT
       alAlaPheThrGlySerProGlnAsnAlaLeuAspPhePheAlaGlnAsnGlyTyrArgC
                        Hpall
       GTCCAGAGGCCTATAATCCGGCCGACTATTTAATAGGTGTACTAGCCTCCGATCCAGGTT
       ysProGluAlaTyrAsnProAlaAspTyrLeuIleGlyValLeuAlaSerAspProGlyT
                         Sau3A
       ATGAAAAGGCTTCCCAAAGATCAGCTCAATATTTGTGTGATCTATTCGCTGTAAGTTCTG
       yrGluLysAlaSerGlnArgSerAlaGlnTyrLeuCysAspLeuPheAlaValSerSerA

       CAGCTAAACAGAGAGACATGTTGGTGAATTTGGAAATACATATGGCTGAAAGTGGTGATT
       laAlaLysGlnArgAspMetLeuValAsnLeuGluIleHisMetAlaGluSerGlyAspT
                           EcoRI
+1000  ATCCTTCTGACAAGGAAGTGGAATTCTTTCGTGCTGCTTCTTTGTATTTAAAATTACATG
       yrProSerAspLysGluValGluPhePheArgAlaAlaSerTrpTyrLeuLysLeuHisV
                                          Sau3A
       TTATCTGGTATAGATACACACTGACACTACTGCGTGATCCTAAACTACAGTGGCTGAGAT
       alIleTrpTyrArgTyrThrLeuThrLeuLeuArgAspProLysLeuGlnTrpLeuArgP
```

```
                                                   HpaII
      TCTTTCAGAAAATGGCCATGGCCATTATAATAGGTGCCTGTTTTGCCGGTACCACGGTAT
      hePheGlnLysMetAlaMetAlaIleIleIleGlyAlaCysPheAlaGlyThrThrValL
        Sau3A
      TGGATCAAATGGGTGTTCAAGCTGTACAGGGTACTCTTTTTGTAATGATTTCTGAAAATA
      euAspGlnMetGlyValGlnAlaValGlnGlyThrLeuPheValMetIleSerGluAsnT

      CTTATCATCCCATGTATTCGGTGTTGAATGTCTTTCCTCAAGGATTTCCATTATTCATGC
      hrTyrHisProMetTyrSerValLeuAsnValPheProGlnGlyPheProLeuPheMetA

+1300 GTGAAACACGTTCTGGCATGTATTCCACAGCTCAGTATTATATTGGCACTGTATTGGCTA
      rgGluThrArgSerGlyMetTyrSerThrAlaGlnTyrTyrIleGlyThrValLeuAlaM

      TGGTAAGATAAGACAAGTAGTAATTAAATGATGATGATGATGACTTGATACATTAATTAA
      et

      AATTCGTATTTGCATGCATGTTCATTGTCTGGTGGCTGGCTGGCTGACTGTTGATTACAT

      GCTATTAATATTTGGTTTTTCCCCCTTCTTTTTATGTTTTCGTTTTGCGCACGTTTATTA
              HpaII
      TCTAGCTGCCGGGCATGATTATAGAACCATTTCTATTTGTTGTCATTTGTTATTTTATCG
           LeuProGlyMetIleIleGluProPheLeuPheValValIleCysTyrPheIleA

+1600 CTGGCTTAAGACCAACATTTTATGCATTTGCTATAACAGCCATAGCTGTTGTACTGGTGA
      laGlyLeuArgProThrPheTyrAlaPheAlaIleThrAlaIleAlaValValLeuValM

      TGAATGTGGCTACAGCATGTGGTTGTTTCTTTTCGACGGCCTTCGATTCGGTACCACTGG
      etAsnValAlaThrAlaCysGlyCysPhePheSerThrAlaPheAspSerValProLeuA
                       HpaII
      CCATGGCATACTTGGTGCCGGTCGATTATATATTCATGATAACATCTGGCATCTTTATTC
      laMetAlaTyrLeuValProValAspTyrIlePheMetIleThrSerGlyIlePheIleG

      AAATCAGGTAAATAAAAAGAAGTATATTTATATGTGACAGTATGTGTTAATGTAGTCTAA
      lnIleSe

      GCTAATTGCAAACCCCCTGAAAATATGCTGTGTTTTAGTACAGTTTTCTGTATATTTCAA
                                                Sau3A
+1900 TGTATGTGTGTTTTTGTGTGCTGTAAACTGTTTAATATCATTGATCTTTGTGTAAAACAC

      TTAATTTTCCTTGTCATCCTGTCGTTGTCGTTTTTATATTCTTTACCTTTGTAGAAAGTG

      CATGTTTGAGTTTGTCATTCCGTTGTAATTTTCCCACCTGTCCGTCTGTCTATCATAGAA
                                                         BamHI
      ATTATGTGTAATTCTTTGATATAATTTCTGATCCTATAAAATATATTTATTTCGGATCCT

      TATAGAAAGCGGAGTTGATTGAGCTATGTCCGTCTGTCTGTCTATCTGTCTGTCTGTCTA

+2200 TCTGTCTGTCTGTCTATCTGTCTATCTGTCTGTCTATCTGTCTGTCTATCTGTCTGTCTA

      TCTGTCTGTCTATCTGTCTGTCTATCTGTCTGTCTATCTGTCTGTCTATCTGTCTGTCTA 300

      ACTGTCTGTCTGTCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTGTCTGTCTATCTG

      TCTGTCTATCTGTCAGCGTGTCTGTCTATCTGTCTGTCCGTCTATCTGTCTGTCTGATTA

      TGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTA
```

```
+2500 TGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTA

      TGTATGTATGTATGTATG (Gap of c. 400 bp)

      TGTCTATCTGTATGTCTGTCTATCTGTCTATTGGATATTTTGGTACTTTTTCGTTCTTTT

      CTTAAAATTAAAGCTAAAAACATGTAACGTATAAAAAGAACTTTTTGTTCTTTAAAATGT

+3100 ACTTTTTGAGTTTCCATAATATTGAACCTAGAAACATGAAATTAGGCATGTAAAGCCCAG

      ATTATGTAAGAACCTATTAAAGGGGTTCCTTTTGGTACTTTTTTGTTCTTAAAAAGGTAT

      TTTTTATTTTTGTTAAACCTAGAAACTGAAATTAAGCATGTAGCCTGCTTAAGGTACGTT
                                                              XbaI
      TTCCTACTCTTGTACTTTTTGAATTTTCTTTAAACTGAGAAACATAAAATTAAACATCTA

      GAGCTAGAAATATGATATTTCCAGATGTACAGCCAAATATAGAACTTTTACTTGTTTAAA

+3400 CACACTATGCTAAAGGGTACTTACGATTCGGAGCAGCCAATTATGTTTTCTTTTTCATTT

      TTGCAAAAAATTCCATTTAATATGTTTTTTTCTCTTATTTTTCAGTACATTGCCTATAGC
                                                   rThrLeuProIleAl
      ATTTTACTGGACACAATTCTTATCATGGATGTTGTATGCAAATGAAGCCATGACTGCTGC
      aPheTyrTrpThrGlnPheLeuSerTrpMetLeuTyrAlaAsnGluAlaMetThrAlaAl

      CCAATGGACTGGTATTACAAATATCAGTAAGTAATTATTTTGTTTTTTTTTTTTGATTGT
      aGlnTrpThrGlyIleThrAsnIleT

      AACACAATTACAGTTTTTCTGTGATTATGTGTATTTGTACAAAGTGGGAATTTACCTTTA

+3700 TGAATACGAGAATTAAACAGCTAAACAGCTAAAGTTTGTTAAATAGATATTAGATTTTAC

      TTATGAAGAATTTAACGGTTATTGTTGAAACATAAGAAA  (Gap of c. 70 bp)

      TATTACCTACACGCAGAAAAAACATGGTTGTGGTAAACATGATATAAGAGCAACATTTTA

      TTGTTAAATTTATGATTTTAGTTTAAACATATTATAGTTATTATTTTACTCAATCATATT

      GAAAGTTATAATTGCCATAGTACAATTAACATTAGTTACAATTAACATAGTATAGTTGTA

+4050 ATAACCATGACTTTTGTCTTAATATTGTTGAAACTTTAATTATATTACGGTGAATCGTAT

      TTTAAGTTATACTTACCGTAAATATTCAAATATTTACAGTTATTTAAATATTTTTAAAAT

      CCAGTTAAAATGACATATTATTTGTTTTTGATAACAGTACATTGCTAAAATGGACATAGT
                             TaqI
      ATGATTGTAGTAAGCATAACTTTTCGACGACGACTGTTAAAACAATAATTATGTTTCGTA

      AATCATATTAAAGCTATAATAGTCATAGACATTAAAATCGTTACAGTAACTAAAATATAT
                                                   TaqI
+4350 GGTTAAAATCCAAATTAGAACAAATAACAGTTACAATGAACATGTTCGAATTAACATAGA

      ATAGTTGTAGTAACCATGACTTTTGACTTAATGACTTTTGCTTCTAAGCCAATAATAAAG

      AATGTTAATTTGTTACTTAATTTACCAAAAAACTCACTTCTTAGTAGTGATAAGGTTTAA

      TTCTCCTTTTAAATAATATGACCCTAAAAAGTTAATTATTTTACTGATATTAATGTTTTC
      CTAGAACATGCAATACATTTTATTTATTACAAAAAATTAATTTAATTGTGTGTATTTCGT

+4650 TTCTTATTTTCTTCCCTTTTGTCTACAGCTTGCTTCGAAGAAAGTGAAAATTTGCCATGT
                                   hrCysPheGluGluSerGluAsnLeuProCys
                          XbaI
      TTTCATACGGGTCAAGATGTTCTAGATAAATACAGTTTTAAAGAAACCAATATCTTTCGT
      PheHisThrGlyGlnAspValLeuAspLysTyrSerPheLysGluThrAsnIlePheArg

      AATCTCTTGGCTATTGTTGGTATTTACTTTGGATTTCATATCTTGGGCTATTATTGCTTG
      AsnLeuLeuAlaIleValGlyIleTyrPheGlyPheHisIleLeuGlyTyrTyrCysLeu

      TGGCGTAGGGCGCGCAAGATATAAGAGAAAAAGTTTCTGCTGCCGAACATTTCTGGTTAT
      TrpArgArgAlaArgLysIleSTP

      ACATATTTATAAAATTTTAAGAACAATTAAAGGATAAAGAATTTAAGTCTATGTGTATTT

+4950 GTGTATTAAATAAATTTACATATGTATTTGTTCGTTTTTTTTTTTGGTACTGATTTATAG
                                   >>>>>>>>>>>>>>>>>
      CAAACATTTATAAAAATACAAAAAAAAAAAATTAACTGAAATAAAATTTTTAACATTAAT
                        <<<<<<<<<<<<<<<<<<
      TTTTATGATATTATTGCGCGCTAAGCAAGGTATGACTATTTGTCTGTCTGTGGCATTGTG

      TCTATTGTGTTTTCAACTTCAAATCAAATAAATTAACAGCCAAAACGTTAGACATTGCAA

      ATGAATAAGAATTTATTTAACACCAAACAAAGAAATTTATTTGCCATCATCAATTTCATG
               XbaI
      CTTCTTTTATCTAGA
```

FIGURE 4.4 The nucleotide sequence and deduced amino acid sequence of the scarlet gene region (modified from Tearle, 1986).

Restriction sites are underlined and marked. Putative promotor, cap site, poly A addition site and palindromes are underlined.

```
                                                          Pst!
-4500                                                 CTGCAGAGAGCAGT

        TAAATTTCCTTACTTAGGTAATAAGTTCAATTAGCAATTGCAGCTTATCAATGATTGTTT

        GCTTTATATATTAGGTGAATATATTTATTATTCATTCATTATTTGTTTTCCTAATTTTGC

        AATAGATTAGATTTATTAAACTACACTTCAGTTGTGGAATATATCACAAAATCGGTTTTT

        TTTTATCTCACTTCTTGATTTTATTAGTTGCGCGTGTATGATTGTATAATTTAACTAATT

-4200   TCAATTAGCGTCTAGTCGGGATTTATGCAGTTTCCATAATCTATTATCAAATCAATTTTT

        TGTTTTTCAATATCAGAAACGTGCAACAAAGTAAAGCACTAAAAACCATCGATAAGGGTT

        TATGAAGTTTAATTAAAGACTCATTCATTTGGATTTATTACACTGGGCAACGATTTGCTC

        CAATAATTATTTTCAGTGATTTGCCGTGCTGTTTGTGTCGTTTGTATAATGAATTTCACA

        TTATTACGCTGCATTATTTTCCATGATATTGTGTTCCAAAGAACATAAATATTTAACCGC

-3600   ACGGATGCGACTGATCGCCATGTAGGTCGATAGCAACTGAAAACCAGAGGGTTGTAGATC

        (Gap of c.500bp)                                         GAT

-3300   CTCATTAAAAATACGTAAATTTACCCAACAAAAAAATACTTCATTGAGAAAAAGCGTCTT

        GTGGAAATGTGAGAATTACAGTTTCTTGTATTTATCATCATTTTCTTGGTTCGCCTTTGA

        GTCAAATAAAATGCGAAAACGCGTAGAAAACAGAGAAGACAATTGCCTCCCATCTGGTTG

        TTGGCTTTCCCGATCGATCCATCGGCGGTTTTCCGTGCTGGCTGCTTGGATGACTCTGGC

          PstI
        GCTGCAGGGCTTCCCGTTGGATTGGTGGGTTGGTCTGGACTGCATTTTCCGCTGGCAATG

-3000   GTAATCGGAGCCGCCACCTAAGCCCGCTGCTTCCCAAGCGGTCGGTAAACTGCAAGACGA

        TCGAATGCCCGCTGCCAGGGCCACCCTAAAGATTGTGGACCCCGCACCTCGGATTCCAAG

        TTTAACATATTTGCATTGTGGGGTGATGAGCACAGTACAGTAGTCATTGGCAAATAAGCT

        ATTGTCATGCAAGCCGCATTATCTTAATAATACGAAAGGAAGAGTGAGAGTAGGGAGTAA

                                        TATA box
        GTAACGAACAAAAATGTAAAATTAATTTTTAAAATAATATGTTATCTATGCAACGATTAA

            Cap
-2700   AAGTATCATTAATTGTAATTAAGAAATATTTTTTTATAACTCTTTAAAAAAAAATAAAAA

        TAAACAAATCCCAGATTCTTTTTGTCAAATCCTCTTTATCCAATCTTTACTGTGCCCGTC

        CATCTTACTGTCAACTTGGCTGTGACAACTTCTCAGCAAAGTTTAGACGGTTAGTTGGTG

        CCCAGCTGTGGCACGCTTCGGTTCAGAAAAAAGTCTGGCTCCTCAGCCGTGAGATCCAAG

        ATAAAAACCAGAGATAATGTCGGATTCAGATAGCAAGCGGATCGATGTGGAGGCCCCGGA
                        MetSerAspSerAspSerLysArgIleAspValGluAlaProGl

-2400   GCGAGTGGAGCAGCACGAATTGCAGGTGATGCCGGTGGGCAGCACCATTGAGGTGCCCAG
        uArgValGluGlnHisGluLeuGlnValMetProValGlySerThrIleGluValProSe

        CTTTGGACAGCACACCCAAGCTATCGAAACGGAACAGTTCCGGAGAGAAGTCTACCGCTC
        rPheGlyGlnHisThrGlnAlaIleGluThrGluGlnPheArgArgGluValTyrArgSe
```

```
          AGGAGCTACAGCAAATGGTCGCCCACGGAGCAGGGAGCCACTCTGGTGTGGCGGGATCTC
          rGlyAlaThrAlaAsnGlyArgProArgSerArgGluProLeuCy

          TGCGTCTATACCAATGTCGGTGGTTCCGGTCAGCGGATGAAGCGGATCATCAACAACTCA

                                                   XhoI
          ACGGGGGCCATTCAACCGGGCACTCTGATGGCTCTAATGGGCTCGAGGTAAGCTGCTGCC

                                        XhoI            XhoI
-2100     GAAATCGATACCACTAAGAGATCAGGTGCACTCGAGTGGAGAAGTGCTCGAGAAGTCGTT

          GTCCAAAAATGTGCTAACTATGTCAAGATGTCTAATTTTCGGCTCTGATAAGCAGAGATT

          CCGCTTGGATATCGCCCCCAACTAATAAAACTTCTACCATTCACAGTGGCTCTGGGAAAA
                                                        sGlySerGlyLysT

          CAACACTCGATGTCAACGCTCGCCTTTCGACAGCCGGGTAAGTTGTGGGCACTACCAAGT
          hrThrLeuMetSerThrLeuAlaPheArgGlnProA

          ACATACGACTCTAGTTCCCGATAAGATTACACTTGATCCCCATACTATCTCCGTCCGCTC

-1800     CATTTGCATTGGGATCTGTGCTGCAGCCGGGACCGTCGTTCAAGGCGACATTCTGATAAA
                                    laGlyThrValValGlnGlyAspIleLeuIleAs

          CGGCCGGCGCATTGGGCCCTTCATGCATCGCAATCACGGCTACGTCTACCAGGATGACCT
          nGlyArgArgIleGlyProPheMetHisArgAsnHisGlyTyrValTyrGlnAspAspLe

                                XhoI
          CTTCCTCGGATCGGTGAGTGTTCTCGAGCACTTGAACTTTATGGTGAGCTACTGTATCTT
          uPheLeuGlySerLeuSerValLeuGluHisLeuAsnPheMet

          AATATCTTAGCTGATAAGTTTATTGAAAGGCTAGAATAACGTTCATTCTTGTTCAAATTA

          ACTATCAAATTGTCTTCTTAAAGTTACCTCAGCTTGTCCTACTTTAAATATATGGTCTAT

-1500     ACACTAGTTTAGTAGGTGGTTAAAAGTATTCTGCTCTCCCTGTTTCAGGCACATCTCCGC
                                                          AlaHisLeuArg

          CTGGATCGTCGGGTGTCGAAAGAGGAGCGTCGCCTCATCATTAAAGAGCTGCTGGAGCGA
          LeuAspArgArgValSerLysGluGluArgArgLeuIleIleLysGluLeuLeuGluArg

          ACCGGCCTTCTTTCGGCGGCGCAAACTCGAATTGGTAGTGGCGATGACAAGAAGGTCCTT
          ThrGlyLeuLeuSerAlaAlaGlnThrArgIleGlySerGlyAspAspLysLysValLeu

          TCGGGGGGAGAACGCAAACGATTGGCATTCGCCGTGGAGCTGCTGAACAATCCGGTGATT
          SerGlyGlyGluArgLysArgLeuAlaPheAlaValGluLeuLeuAsnAsnProValIle

          CTATTTTGCGATGAGCCTACCACGGGACTGGACTCATACAGTGCCCAGCAGCTGGTGGCC
          LeuPheCysAspGluProThrThrGlyLeuAspSerTyrSerAlaGlnGlnLeuValAla

-1200     ACGTTGTACGAGTTGGCCCAAAAGGGCACCACCATACTGTGCACCATCCATCAGCCGAGT
          ThrLeuTyrGluLeuAlaGlnLysGlyThrThrIleLeuCysThrIleHisGlnProSer

          TCGCAGCTCTTCGATAACTTTAACAACGTAATGTTGCTGGCCGATGGGCGAGTAGCCTTT
          SerGlnLeuPheAspAsnPheAsnAsnValMetLeuLeuAlaAspGlyArgValAlaPhe

          ACGGGATCACCTCAGCATGCGCTTAGTTTCTTTGCGAATCATGGATACTACTGCCCGGAG
          ThrGlySerProGlnHisAlaLeuSerPhePheAlaAsnHisGlyTyrTyrCysProGlu

          GCCTACAATCCGGCAGATTTCCTAATTGGTGTCCTGGCGACCGATCCTGGCTATGAGCAG
          AlaTyrAsnProAlaAspPheLeuIleGlyValLeuAlaThrAspProGlyTyrGluGln
```

```
       GCCTCGCAGAGATCGGCTCAACACCTTTGTGATCAGTTTGCCGTCAGCTCGGCGGCCAAG
       AlaSerGlnArgSerAlaGlnHisLeuCysAspGlnPheAlaValSerSerAlaAlaLys

-900   CAGCGGGATATGCTGGTTAATCTGGAGATTCACATGGCCCAGTCAGGCAACTTTCCCTTC
       GlnArgAspMetLeuValAsnLeuGluIleHisMetAlaGlnSerGlyAsnPheProPhe

       GACACGGAGGTGGAGTCCTTCAGGGGCGTGGCGTGGTACAAGCGCTTCCACGTAGTGTGG
       AspThrGluValGluSerPheArgGlyValAlaTrpTyrLysArgPheHisValValTrp

                                 BamHI
       CTAAGGGCGATCGTGACGCTGCTAAGGGATCCCACGATTCAATGGTTGCGGTTCATTCAA
       LeuArgAlaIleValThrLeuLeuArgAspProThrIleGlnTrpLeuArgPheIleGln

       AAGATCGCAATGGCATTTATTATCGGTGCCTGTTTTGCCGGAACCACGGAACCCTCCCAGT
       LysIleAlaMetAlaPheIleIleGlyAlaCysPheAlaGlyThrThrGluProSerGlnL

       TGGGCGTACAGGCTGTTCAGGGAGCACTTTTCATTATGATATCGGAGAACACCTACCATC
       euGlyValGlnAlaValGlnGlyAlaLeuPheIleMetIleSerGluAsnThrTyrHisP

-600   CCATGTACTCCGTGCTGAATCTCTTCCCGCAGGGATTTCCGCTATTCATGCGGGAAACCC
       roMetTyrSerValLeuAsnLeuPheProGlnGlyPheProLeuPheMetArgGluThrA

       GATCCGGACTCTACTCCACGGGACAATATTATGCGGCCAATATACTGGCTTTGGTAAGTG
       rgSerGlyLeuTyrSerThrGlyGlnTyrTyrAlaAlaAsnIleLeuAlaLeu

       TTGTAACGAACTAATTGGATTTAATTTAAAAAGGAGCATCATTTTAAAGTTTTCACATGT

       CTGTTAATAGCTGCCTGGCATGATAATTGAGCCCCTGATATTCGTCATAATCTGCTACTG
                 LeuProGlyMetIleIleGluProLeuIlePheValIleIleCysTyrTr

       GCTGACGGGTCTGAGATCCACCTTTTATGCCTTCGGAGTGACTGCCATGTGTGTGGTGCT
       pLeuThrGlyLeuArgSerThrPheTyrAlaPheGlyValThrAlaMetCysValValLe

-300   GGTGATGAATGTGGCCACAGCCTGCGGTTGCTTCTTTTCCACGGCCTTTAATTCGGTGCC
       uValMetAsnValAlaThrAlaCysGlyCysPhePheSerThrAlaPheAsnSerValPr

       GCTGGCAATGGCTTACTTGGTGCCCTTGGATTATATATTCATGATCACCTCGGGAATCTT
       oLeuAlaMetAlaTyrLeuValProLeuAspTyrIlePheMetIleThrSerGlyIlePh

       TATACAAGTGAAGTGGGTGATTAAATTGATACTTTCAATTAAGTTCGTTTCAAGAAACAC
       eIleGlnValAs

       TCAACTTTCAGTTCGCTACCAGTGGCGTTTTGGTGGACACAATTCCTCTCATGGATGCTG
                  nSerLeuProValAlaPheTrpTrpThrGlnPheLeuSerTrpMetLeu

       TATGCCAATGAGGCAATGACGGCTGCTCAATGGTCTGGAGTGCAGAATATAAGTAAGTTT    -1
       TyrAlaAsnGluAlaMetThrAlaAlaGlnTrpSerGlyValGlnAsnIleT

       HindIII
+1     AAGCTTTAATACCCTGTTCCAAGATTAATATTAATTGAAATTTGAATAGCCTGTTTTCAG
                                                        hrCysPheGln

       GAGAGTGCCGACTTGCCGTGCTTTCACACGGGTCAGGATGTCCTGGACAAGTACACCTTC
       GluSerAlaAspLeuProCysPheHisThrGlyGlnAspValLeuAspLysTyrThrPhe

       AACGAGAGCAATGTCTATCGGAATTTACTGGCCATGGTGGGTCTTTATTTCGGATTCCAT
       AsnGluSerAsnValTyrArgAsnLeuLeuAlaMetValGlyLeuTyrPheGlyPheHis

       CTACTGGGATATTATTGCCTTTGGCGAAGGGCGCGGAAGCTGTAAATATACTTACTACGG
       LeuLeuGlyTyrTyrCysLeuTrpArgArgAlaArgLysLeuSTP
```

```
AGCTATGAACTCATCAAATAAACAACAATGCCTAGAATTTTCAAATTAATAGCGCCATTT   300
   >>>>>>>>>>>>>>>>>>>>>>>>>
TTCAACACACTTTGTTTTTTATTTCCTTTGTTTTTAGAGCACAATGAAACACATTTAGTT
           <<<<<<<<<<<<<<<<<<<<<<<<<

AGACTGGGCTTAGTTTTGTCCTTACCAACTTGAAACTGAACATCAAAGCTAATCAACAAC

GAGTTCAATACACGCATACACTTATGATTATATACATAACATATAGGATACTTTATACTC

AGTTAAATTATAGTTCGTAAGTAGCTAGGCGTAGTCAAAGTGGC(Gap of c. 250bp)

TCAAATTTAAACAATGCTATTTCAAGACATAACGTATTAGATACTCTCTTCATCTCAACT

TATACTCTAGTAAATATCGTTTGTGCAACAATTCAAAAAGTTTCCATGTGCTCTTCAATT   900

GACAGATTTGGTTTTAGTTGTTAAGAAAACACAGCTATGAACAGAGTCAAAAACCGATGT

CATATTTCTTATCGATATGGTATGTAATGTATGCTGTCCCTTAGTTTGGTGTAAAAATAT

TTTGCTGGGGAATGCAATCGAGTCATAAAACATTCAATTTGAGACGGAAAAATATTTCCT

TTTCATAAAATTAGAACTGCATGTTATGGATTTCATTAATGAAAAATACATTGAAGAAT

GACCAGAAATCATAGATTATATATTATATTTTTATATGTTCATATACGAAGTACAACGCC  1200

ATTTATTGTATTCCCCAACATCAAGTAATAACTAATTATTCGTTTGACAATGAACATTAA

ATTCGATTTCTTTTGCCCGCGTATTCCCTTTGAACTTACACTCTACTGCAAACACACATA

GGAACCTTCACACCTACATACCTATACTCCTTTGCAAGGACTCTGCCCCAAACTTACAAC

CTATCGAATGCCTAGCAGTTCTAACAATGCTCTTTGGTACCCAGTTGGAGTGCAACACTT

                                     EcoRI
TGCTTTTCGTACATTAGTTTTATATGTTGTGATATGGAATTC                    1500
```

TABLE 4.1: Topaz EXON/INTRON JUNCTION SEQUENCES

	DONOR SEQUENCE	ACCEPTOR SEQUENCE
	exon/intron	intron/exon
consensus (Mount, 1982)	$\frac{C}{A}$AG/GT$\frac{A}{G}$AGT	$\frac{T}{C}_n$ N$\frac{T}{C}$AG/G (n≥15)
exon/intron 2	CAG/GTAATT	ATTTCCAG/T
exon/intron 3	ATG/GTAGGA	TATTTTAG/G
exon/intron 4	ATG/GTAAGA	TTATCTAG/C
exon/intron 5	CAG/GTAAAT	TTTTTCAG/T
exon/intron 6	TCA/GTAAGT	GTCTACAG/C

TABLE 4.2: THE TOPAZ AND SCARLET EXONS

			% Homology			% GC	
Exon	Length in bp	Number of amino acids	DNA level	Predicted amino acid level	% Conservative amino acid substitutions/ total number of substitutions	topaz	scarlet
2	49	16	69	87	50	41	57
3	137	46	67	80	22	37	55
4	966	322	68	81	52	40	54
5	242	81	75	83	86	41	50
6	101	33	74	82	100	38	49
7	173	58	73	81	91	36	51

detected, no homology to the equivalent region in scarlet was observed, therefore the position and structure of the first exon could not be established. It was initially thought either that the DNA sequence of the first exons might have diverged at a higher rate between the two genes than the other exons and therefore no homology could be detected, or that the first exon might be very short and would not be recognized by comparing sequences from the two genes. However recently, Northern analysis of D.melanogaster RNA using fragments from along the scarlet gene region as probes, has been carried out (F.V. Morris and A.J. Howells, unpublished). This has shown that all fragments that carry putative scarlet exons hybridize to a 2.3 kb polyA$^+$ transcript. In addition, the 0.9 kb PstI/XhoI fragment (-3.2 to -2.3) also hybridizes to this scarlet transcript, but the next upstream fragment (1.5 kb PstI) hybridizes to a different size transcript, i.e. is part of a different transcriptional unit. This analysis strongly indicated that the first exon is contained within the 0.9 kb PstI/XhoI fragment. Careful analysis of the sequence of this region revealed seven open reading frames which begin with an ATG, (ranging in size from 42 to 210 bp) and all are followed by an intron donor site, creating a two bp overhang that could be spliced to join the one bp overhang in exon 2, thereby maintaining the reading-frame. The longest potential first exon codes for 70 amino acids and is presented in Figure 4.4.

As a consequence of the above analysis, hybridizations of subclone pstP/XhO.9 to the topaz phage DNA under low stringency conditions (50^{o}C, 3xSSC) were carried out and a weak hybridization signal to the 3.7 kb EcoR1/SalI fragment in λto8 (-7.8 to -4.1) was found. Further studies to precisely define the region of homology have been hampered by the lack of recognition sites for the commonly used restriction enzymes within the subclone ptoES3.7 (see Figure 3.7) and the region of homology to the putative scarlet first exon has been narrowed down only to a 2.6 kb SalI/XhoI fragment (Figure 4.5). Analysis of 300 bp of sequence (derived by sequencing from the XhoI site towards the SalI site (-5100 to -5400 Figure 4.3)) did not reveal homology to any sequences in the pstP/Xh0.9 subclone, indicating the region of homology is further upstream from the XhoI site.*

4.2.2 The exons

The homology between topaz and scarlet sequences is restricted to the putative exon regions. The exons are identical in size in the two genes and are about 70% homologous at the nucleotide level and 80% homologous at the predicted amino acid level (Figure 4.6 and Table 4.2). A striking difference between the exons from the two genes is in their GC composition and their codon preference (Tables 4.2 and 4.3). Topaz has a much lower GC composition than scarlet. This is reflected in the codon preference and despite a very similar amino acid composition (Table 4.4) topaz has only 32% GC in third base position compared to 64% GC in scarlet.

4.2.3 The introns

The putative topaz and scarlet introns are located at identical positions in the two genes but share no homology in sequence or (with one

* These recent experiments to locate exon 1 were carried out while the preparation of the thesis was in progress and I would like to acknowledge the technical assistance of Mrs V. Corrigan in the hybridization experiments involving the first exon.

FIGURE 4.5 Putative localization of the first exon in *topaz*.

Restriction map of the *topaz* and *scarlet* gene regions. The *scarlet* subclone pstP/XhO.9 hybridizes to the region marked in ▥

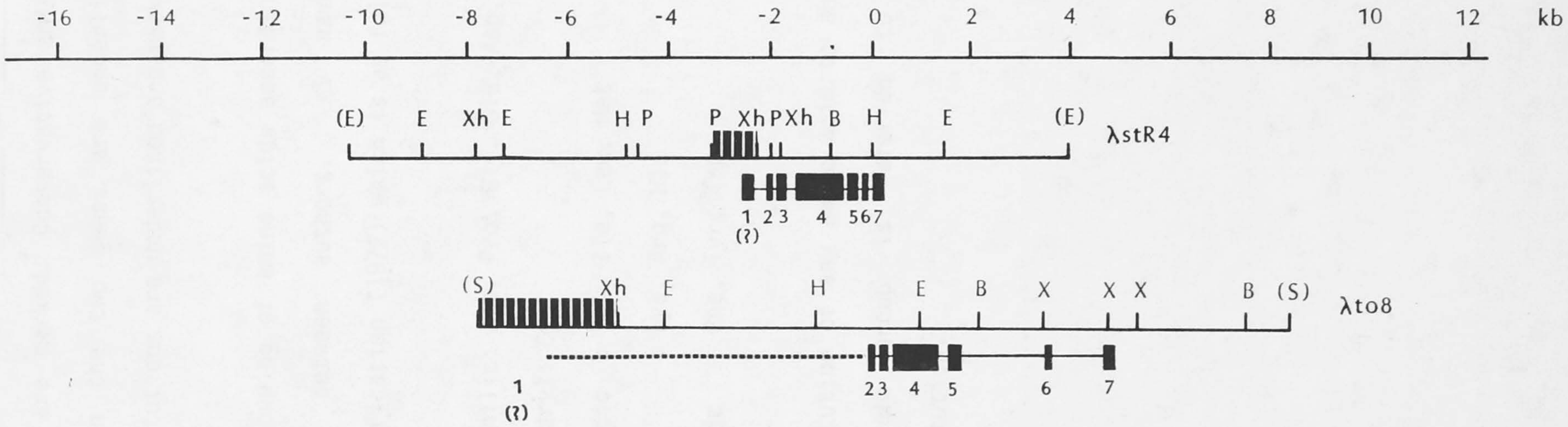

-16
-14
-12
-10
-8
-6
-4
-2
0
2
4
6
8
10
12
kb
(E)
E
Xh
E
H
P
P
Xh
P
Xh
B
H
E
(E)
λstR4
1
(?)
2
3
4
5
6
7
(S)
Xh
E
H
E
B
X
X
X
B
(S)
λto8
1
(?)
2
3
4
5
6
7

FIGURE 4.6 The nucleotide sequence and predicted amino acid sequence of exons 2 to 7 from topaz and scarlet. The position of the exons are marked. Conservative amino acid substitutions between the two genes are underlined. Non-conservative substitutions are underlined and marked with an asterisk *.

The grouping of amino acids according to their properties varies between authors. I have used the Dayhoff classification (1972) which is as follows:

Hydrophilic: Ala, Pro, Gly, Glu, Asp, Gln, Asn, Ser, Thr

Sulphydryl: Cys

Aliphatic : Val, Ile, Leu, Met

Basic : Lys, Arg, His

Aromatic : Phe, Tyr, Trp.

Substitution of any amino acid to another amino acid from the same group is referred to as a conservative substitution.

```
|Exon 2                                          |Exon 3
                            *                    *
GGTTCAGGAAAAACAACATTAATGTCTGTACTTGCATATCGCCAACCAGTTGGTACCGTA to
GlySerGlyLysThrThrLeuMetSerValLeuAlaTyrArgGlnProValGlyThrVal
GGCTCTGGGAAAACAACACTCATGTCAACGCTCGCCTTTCGACAGCCGGCCGGGACCGTC st
GlySerGlyLysThrThrLeuMetSerThrLeuAlaPheArgGlnProAlaGlyThrVal

                                                       *  *
GTTCAAGGTGATATTCTTATCAATGGCCGACGTATAGGACCATTTATGCATCGCATAAGT
ValGlnGlyAspIleLeuIleAsnGlyArgArgIleGlyProPheMetHisArgIleSer
GTTCAAGGCGACATTCTGATAAACGGCCGGCGCATTGGGCCCTTCATGCATCGCAATCAC
ValGlnGlyAspIleLeuIleAsnGlyArgArgIleGlyProPheMetHisArgAsnHis

    *                       *                 *           *
GGTTGTGTTTATCAAGATGATTTATTTAATGGATCACTTACCGTGGCAGAACATATGCAC
GlyCysValTyrGlnAspAspLeuPheAsnGlySerLeuThrValAlaGluHisMetHis
GGCTACGTCTACCAGGATGACCTCTTCCTCGGATCGCTGAGTGTTCTCGAGCACTTGAAC
GlyTyrValTyrGlnAspAspLeuPheLeuGlySerLeuSerValLeuGluHisLeuAsn

     |Exon4
          *
TTTATGGCTCTCTTACGTTTAGATCGCCGCGTTAGCAAGCAGGAACGTAAACTTATAATA
PheMetAlaLeuLeuArgLeuAspArgArgValSerLysGlnGluArgLysLeuIleIle
TTTATGGCACATCTCCGCCTGGATCGTCGGGTGTCGAAAGAGGAGCGTCGCCTCATCATT
PheMetAlaHisLeuArgLeuAspArgArgValSerLysGluGluArgArgLeuIleIle

 *        *
CAAGATCTTTTCGAACGTACAGGTCTATTGGGTGCTTCTAATACACGTATTGGTTCGGGA
GlnAspLeuPheGluArgThrGlyLeuLeuGlyAlaSerAsnThrArgIleGlySerGly
AAAGAGCTGCTGGAGCGAACCGGCCTTCTTTCGGCGGCGCAAACTCGAATTGGTAGTGGC
LysGluLeuLeuGluArgThrGlyLeuLeuSerAlaAlaGlnThrArgIleGlySerGly

       *
GATGATGAAAAAGTGTTATCGGGTGGTGAACGTAAACGTTTAGCTTTTGCTGTGGAATTG
AspAspGluLysValLeuSerGlyGlyGluArgLysArgLeuAlaPheAlaValGluLeu
GATGACAAGAAGGTCCTTTCGGGGGGAGAACGCAAACGATTGGCATTCGCCGTGGAGCTG
AspAspLysLysValLeuSerGlyGlyGluArgLysArgLeuAlaPheAlaValGluLeu

TTAAATAATCCGGTGATATTATTTTGTGATGAACCCACCACTGGTTTGGATTCTTATAGT
LeuAsnAsnProValIleLeuPheCysAspGluProThrThrGlyLeuAspSerTyrSer
CTGAACAATCCGGTGATTCTATTTTGCGATGAGCCTACCACGGGACTGGACTCATACAGT
LeuAsnAsnProValIleLeuPheCysAspGluProThrThrGlyLeuAspSerTyrSer

                                     *
GCTCAGCAGTTGGTGCAAACCCTTTACGATTTAGCCAAAAAGGGTACCACTATCTTATGC
AlaGlnGlnLeuValGlnThrLeuTyrAspLeuAlaLysLysGlyThrThrIleLeuCys
GCCCAGCAGCTGGTGGCCACGTTGTACGAGTTGGCCCAAAAGGGCACCACCATACTGTGC
AlaGlnGlnLeuValAlaThrLeuTyrGluLeuAlaGlnLysGlyThrThrIleLeuCys

                                  *                 *
ACCATACACCAACCGTCTTCACAATTATTTGATATGTTTAATAATGTTCTCTTTTTGTCG
ThrIleHisGlnProSerSerGlnLeuPheAspMetPheAsnAsnValLeuPheLeuSer
ACCATCCATCAGCCGAGTTCGCAGCTCTTCGATAACTTTAACAACGTAATGTTGCTGGCC
ThrIleHisGlnProSerSerGlnLeuPheAspAsnPheAsnAsnValMetLeuLeuAla

                                  *                       *
GAGGGCAGAGTGGCCTTTACTGGTTCACCACAAAATGCTTTGGATTTTTTTGCTCAAAAT
GluGlyArgValAlaPheThrGlySerProGlnAsnAlaLeuAspPhePheAlaGlnAsn
GATGGGCGAGTAGCCTTTACGGGATCACCTCAGCATGCGCTTAGTTTCTTTGCGAATCAT
AspGlyArgValAlaPheThrGlySerProGlnHisAlaLeuSerPhePheAlaAsnHis
```

```
       *
GGTTATAGATGTCCAGAGGCCTATAATCCGGCCGACTATTTAATAGGTGTACTAGCCTCC to
GlyTyrArgCysProGluAlaTyrAsnProAlaAspTyrLeuIleGlyValLeuAlaSer
GGATACTACTGCCCGGAGGCCTACAATCCGGCAGATTTCCTAATTGGTGTCCTGGCGACC st
GlyTyrTyrCysProGluAlaTyrAsnProAlaAspPheLeuIleGlyValLeuAlaThr

                *                       *           *
GATCCAGGTTATGAAAAGGCTTCCCAAAGATCAGCTCAATATTTGTGTGATCTATTCGCT
AspProGlyTyrGluLysAlaSerGlnArgSerAlaGlnTyrLeuCysAspLeuPheAla
GATCCTGGCTATGAGCAGGCCTCGCAGAGATCGGCTCAACACCTTTGTGATCAGTTTGCC
AspProGlyTyrGluGlnAlaSerGlnArgSerAlaGlnHisLeuCysAspGlnPheAla

GTAAGTTCTGCAGCTAAACAGAGAGACATGTTGGTGAATTTGGAAATACATATGGCTGAA
ValSerSerAlaAlaLysGlnArgAspMetLeuValAsnLeuGluIleHisMetAlaGlu
GTCAGCTCGGCGGCCAAGCAGCGGGATATGCTGGTTAATCTGGAGATTCACATGGCCCAG
ValSerSerAlaAlaLysGlnArgAspMetLeuValAsnLeuGluIleHisMetAlaGln

                *     *           *           *           *
AGTGGTGATTATCCTTCTGACAAGGAAGTGGAATTCTTTCGTGCTGCTTCTTGGTATTTA
SerGlyAspTyrProSerAspLysGluValGluPhePheArgAlaAlaSerTrpTyrLeu
TCAGGCAACTTTCCCTTCGACACGGAGGTGGAGTCCTTCAGGGGCGTGGCGTGGTACAAG
SerGlyAsnPheProPheAspThrGluValGluSerPheArgGlyValAlaTrpTyrLys

    *              *     *  *                       *
AAATTACATGTTATCTGGTATAGATACACACTGACACTACTGCGTGATCCTAAACTACAG
LysLeuHisValIleTrpTyrArgTyrThrLeuThrLeuLeuArgAspProLysLeuGln
CGCTTCCACGATGTGTGGCTAAGGGCGATCGTGACGCTGCTAAGGGATCCCACGATTCAA
ArgPheHisValValTrpLeuArgAlaIleValThrLeuLeuArgAspProThrIleGln

             *                    *
TGGCTGAGATTCTTTCAGAAAATGGCCATGGCCATTATAATAGGTGCCTGTTTTGCCGGT
TrpLeuArgPhePheGlnLysMetAlaMetAlaIleIleIleGlyAlaCysPheAlaGly
TGGTTGCGGTTCATTCAAAAGATCGCAATGGCATTTATTATCGGTGCCTGTTTTGCCGGA
TrpLeuArgPheIleGlnLysIleAlaMetAlaPheIleIleGlyAlaCysPheAlaGly

       *  *
ACCACGGTATTGGATCAAATGGGTGTTCAAGCTGTACAGGGTACTCTTTTTGTAATGATT
ThrThrValLeuAspGlnMetGlyValGlnAlaValGlnGlyThrLeuPheValMetIle
ACCACGGAACCCTCCCAGTTGGGCGTACAGGCTGTTCAGGGAGCACTTTTCATTATGATA
ThrThrGluProSerGlnLeuGlyValGlnAlaValGlnGlyAlaLeuPheIleMetIle

TCTGAAAATACTTATCATCCCATGTATTCGGTGTTGAATGTCTTTCCTCAAGGATTTCCA
SerGluAsnThrTyrHisProMetTyrSerValLeuAsnValPheProGlnGlyPhePro
TCGGAGAACACCTACCATCCCATGTACTCCGTGCTGAATCTCTTCCCGCAGGGATTTCCG
SerGluAsnThrTyrHisProMetTyrSerValLeuAsnLeuPheProGlnGlyPhePro

                                                    *
TTATTCATGCGTGAAACACGTTCTGGCATGTATTCCACAGCTCAGTATTATATTGGCACT
LeuPheMetArgGluThrArgSerGlyMetTyrSerThrAlaGlnTyrTyrIleGlyThr
CTATTCATGCGGGAAACCCGATCCGGACTCTACTCCACGGGACAATATTATGCGGCCAAT
LeuPheMetArgGluThrArgSerGlyLeuTyrSerThrGlyGlnTyrTyrAlaAlaAsn

           |Exon 5
                                     *
GTATTGGCTATGCTGCCGGGCATGATTATAGAACCATTTCTATTTGTTGTCATTTGTTAT
ValLeuAlaMetLeuProGlyMetIleIleGluProPheLeuPheValValIleCysTyr
ATACTGGCTTTGCTGCCTGGCATGATAATTGAGCCCCTGATATTCGTCATAATCTGCTAC
IleLeuAlaLeuLeuProGlyMetIleIleGluProLeuIlePheValIleIleCysTyr
```

```
                                                     *
TTTATCGCTGGCTTAAGACCAACATTTTATGCATTTGCTATAACAGCCATAGCTGTTGTA to
PheIleAlaGlyLeuArgProThrPheTyrAlaPheAlaIleThrAlaIleAlaValVal
TGGCTGACGGGTCTGAGATCCACCTTTTATGCCTTCGGAGTGACTGCCATGTGTGTGGTG st
TrpLeuThrGlyLeuArgSerThrPheTyrAlaPheGlyValThrAlaMetCysValVal

CTGGTGATGAATGTGGCTACAGCATGTGGTTGTTTCTTTTCGACGGCCTTCGATTCGGTA
LeuValMetAsnValAlaThrAlaCysGlyCysPhePheSerThrAlaPheAspSerVal
CTGGTGATGAATGTGGCCACAGCCTGCGGTTGCTTCTTTTCCACGGCCTTTAATTCGGTG
LeuValMetAsnValAlaThrAlaCysGlyCysPhePheSerThrAlaPheAsnSerVal

CCACTGGCCATGGCATACTTGGTGCCGGTCGATTATATATTCATGATAACATCTGGCATC
ProLeuAlaMetAlaTyrLeuValProValAspTyrIlePheMetIleThrSerGlyIle
CCGCTGGCAATGGCTTACTTGGTGCCCTTGGATTATATATTCATGATCACCTCGGGAATC
ProLeuAlaMetAlaTyrLeuValProLeuAspTyrIlePheMetIleThrSerGlyIle

              |Exon 6

TTTATTCAAATCAGTACATTGCCTATAGCATTTTACTGGACACAATTCTTATCATGGATG
PheIleGlnIleSerThrLeuProIleAlaPheTyrTrpThrGlnPheLeuSerTrpMet
TTTATACAAGTGAATTCGCTACCAGTGGCGTTTTGGTGGACACAATTCCTCTCATGGATG
PheIleGlnValAsnSerLeuProValAlaPheTrpTrpThrGlnPheLeuSerTrpMet

                                                       |Exon 7

TTGTATGCAAATGAAGCCATGACTGCTGCCCAATGGACTGGTATTACAAATATCACTTGC
LeuTyrAlaAsnGluAlaMetThrAlaAlaGlnTrpThrGlyIleThrAsnIleThrCys
CTGTATGCCAATGAGGCAATGACGGCTGCTCAATGGTCTGGAGTGCAGAATATAACCTGT
LeuTyrAlaAsnGluAlaMetThrAlaAlaGlnTrpSerGlyValGlnAsnIleThrCys

TTCGAAGAAAGTGAAAATTTGCCATGTTTTCATACGGGTCAAGATGTTCTAGATAAATAC
PheGluGluSerGluAsnLeuProCysPheHisThrGlyGlnAspValLeuAspLysTyr
TTTCAGGAGAGTGCCGACTTGCCGTGCTTTCACACGGGTCAGGATGTCCTGGACAAGTAC
PheGlnGluSerAlaAspLeuProCysPheHisThrGlyGlnAspValLeuAspLysTyr

AGTTTTAAAGAAACCAATATCTTTCGTAATCTCTTGGCTATTGTTGGTATTTACTTTGGA
SerPheLysGluThrAsnIlePheArgAsnLeuLeuAlaIleValGlyIleTyrPheGly
ACCTTCAACGAGAGCAATGTCTATCGGAATTTACTGGCCATGGTGGGTCTTTATTTCGGA
ThrPheAsnGluSerAsnValTyrArgAsnLeuLeuAlaMetValGlyLeuTyrPheGly

TTTCATATCTTGGGCTATTATTGCTTGTGGCGTAGGGCGCGCAAGATA
PheHisIleLeuGlyTyrTyrCysLeuTrpArgArgAlaArgLysIle
TTCCATCTACTGGGATATTATTGCCTTTGGCGAAGGGCGCGGAAGCTG
PheHisLeuLeuGlyTyrTyrCysLeuTrpArgArgAlaArgLysLeu
```

TABLE 4.3: CODON PREFERENCE FOR THE TOPAZ AND SCARLET GENES

TOPAZ

AAA (Lys): 11	AAG (Lys): 5	AAC (Asn); 0	AAT (Asn): 19
ACA (Thr): 15	ACG (Thr): 3	AAC (Thr): 8	ACT (Thr): 9
AGA (Arg): 7	AGG (Arg): 1	AGC (Ser): 1	AGT (Ser): 7
ATA (Ile): 17	ATG (Met): 21	ATC (Ile): 9	ATT (Ile): 11
CAA (Gln): 18	CAG (Gln): 8	CAC (His): 2	CAT (His): 7
CCA (Pro): 10	CCG (Pro): 5	CCC (Pro): 2	CCT (Pro): 4
CGA (Arg): 1	CGG (Arg): 0	CGC (Arg): 5	CGT (Arg): 13
CTA (Leu): 7	CTG (Leu): 6	CTC (Leu): 3	CTT (Leu): 7
GAA (Glu): 20	GAG (Glu): 2	GAC (Asp): 3	GAT (Asp): 21
GCA (Ala): 8	GCG (Ala): 1	GCC (Ala): 14	GCT (Ala): 23
GGA (Gly): 6	GGG (Gly): 0	GGC (Gly): 8	GGT (Gly): 24
GTA (Val): 10	GTG (Val): 12	GTC (Val): 3	GTT (Val): 11
TAA (STOP): 1	TAG (STOP): 0	TAC (Tyr): 6	TAT (Tyr): 22
TCA (Ser): 6	TCG (Ser): 6	TCC (Ser): 3	TCT (Ser): 10
TGA (STOP): 0	TGG (Trp): 7	TGC (Cys): 3	TGT (Cys): 9
TTA (Leu): 17	TTG (Leu): 19	TCC (Phe): 10	TTT (Phe): 30

32% GC at third base position

SCARLET

AAA (Lys): 4	AAG (Lys): 8	AAC (Asn): 9	AAT (Asn): 14
ACA (Thr): 4	ACG (Thr): 13	ACC (Thr): 14	ACT (Thr): 2
AGA (Arg): 2	AGG (Arg): 4	AGC (Ser): 2	AGT (Ser): 6
ATA (Ile): 10	ATG (Met): 18	ATC (Ile): 8	ATT (Ile): 12
CAA (Gln): 10	CAG (Gln): 19	CAC (His): 6	CAT (His): 7
CCA (Pro): 1	CCG (Pro): 9	CCC (Pro): 7	CCT (Pro): 4
CGA (Arg): 7	CGG (Arg): 7	CGC (Arg): 6	CGT (Arg): 2
CTA (Leu): 7	CTG (Leu): 28	CTC (Leu): 11	CTT (Leu): 8
GAA (Glu): 3	GAG (Glu): 18	GAC (Asp): 7	GAT (Asp): 14
GCA (Ala): 8	GCG (Ala): 11	GCC (Ala): 24	GCT (Ala): 6
GGA (Gly): 15	GGG (Gly): 5	GGC (Gly): 12	GGT (Gly): 7
GTA (Val): 3	GTG (Val): 20	GTC (Val): 8	GTT (Val): 4
TAA (STOP): 1	TAG (STOP): 0	TAC (Tyr): 14	TAT (Tyr): 10
TCA (Ser): 5	TCG (Ser): 12	TCC (Ser): 7	TCT (Ser): 2
TGA (STOP): 0	TGG (Trp): 9	TGC (Cys): 8	TGT (Cys): 4
TTA (Leu): 1	TTG (Leu): 11	TTC (Phe): 21	TTT (Phe): 18

65% GC at third base position

TABLE 4.4: THE AMINO ACID COMPOSITION OF THE TOPAZ AND SCARLET GENES (EXON 2 TO 7)

TOPAZ

Ala: 46 (8.3%)	Arg: 27 (4.9%)	Asn: 19 (3.4%)	Asp: 24 (4.3%)
Cys: 12 (2.2%)	Gln: 26 (4.7%)	Glu: 22 (4.0%)	Gly: 38 (6.8%)
His: 9 (1.6%)	Ile: 37 (6.7%)	Leu: 59 (10.6%)	Lys: 16 (2.9%)
Met: 21 (3.8%)	Phe: 40 (7.2%)	Pro: 21 (3.8%)	Ser: 33 (5.9%)
Thr: 35 (6.3%)	Trp: 7 (1.3%)	Tyr: 28 (5.0%)	Val: 36 (6.5%)

SCARLET

Ala: 49 (8.8%)	Arg: 28 (5.0%)	Asn: 23 (4.1%)	Asp: 21(3.8%)
Cys: 12 (2.2%)	Gln: 29 (5.2%)	Glu: 21 (3.8%)	Gly: 39 (7.0%)
His: 13 (2.3%)	Ile: 30 (5.4%)	Leu: 66 (11.9%)	Lys: 12 (2.2%)
Met: 18 (3.2%)	Phe: 39 (7.0%)	Pro: 21 (3.8%)	Ser: 34 (6.1%)
Thr: 33 (5.9%)	Trp: 9 (1.6%)	Tyr: 24 (4.3%)	Val: 35 (6.3%)

exception), in size (Table 4.5). Intron 1, although not fully defined, is probably longer than 6 kb in topaz while it is only a few hundred base pairs in scarlet. Intron 2 is the only intron which is identical in size in the two genes, while introns 3 and 4 in topaz are similar in size to introns 4 and 3 in scarlet, respectively. Introns 5 and 6 in topaz are considerably longer than the same introns in scarlet, with the overall affect being that the topaz gene is about 4 times longer than the scarlet gene. The GC composition of the topaz introns is lower than the scarlet introns (Table 4.5) and in both genes is lower than that observed in the exons.

4.2.4 Repeated sequences

The presence of repeated sequences within the topaz gene region was identified first by hybridization (Chapter 3) and in this chapter was confirmed by DNA sequencing. The repeated sequences are located in introns 1,5 and 6. When judged by hybridization intensity (using nick-translated genomic DNA as the probe), the 3.2 EcoR1/HindIII fragment of intron 1 appears to contain sequences which are highly repeated in the L.cuprina genome. Only ~600 bp of this fragment was sequenced (partly because no restriction sites convenient for subcloning were found that would cut within the fragment). The 300 bp of sequence coming downstream from the EcoR1 site (-4100) contains a repeated sequence with an imperfect repeat unit 12bp long (see Figure 4.3 and Table 4.6); the sequence going upstream from the HindIII site (-840) does not show any internal repetition. Whether the 12 bp repeat is the only repeated sequence contained within this fragment or whether there are other families of repeated DNA in it as well is unknown and could only be established by further characterization. The

TABLE 4.5: THE INTRONS OF THE <u>TOPAZ</u> AND <u>SCARLET</u> GENES

	Topaz	Scarlet
intron: length in bp		
1	∿6000(?)	300(?)
2	109	109
3	55	185
4	183	77
5	∿1700	59
6	∿1050	57
Total % GC	27	38

TABLE 4.6: REPEATED SEQUENCES IN THE TOPAZ GENE

Position (in Fig. 4.3)	Intron	Repeat unit in bp	Most Common sequence	Comments
-4100 to -3860	1	12	TATNGN$\frac{C}{T}$TAGTC$\frac{T}{C}$	Sequence variations between repeats, organized in almost perfect tandem array
2172 to 2578, 2980 to 3010	5	4	TCT$\frac{G}{A}$, TGTA	Almost perfect tandem array
3024 to 4400	5,6	13 or 14	ACTTTT(T)(C)GTTCTT	Dispersed repeat; (C) or (T) sometimes missing
3993 to 4427	6	15	AGTT$\frac{G}{A}$AATN$\frac{G}{A}$ACAT	Some repeats organized in tandem, some are dispersed among other sequences.

Examples of each type of repeat unit are underlined in Figure 4.3.

EcoR1/SalI fragment (-7.8 to -4.1) that probably carries exon 1 also contains repeated sequences (as judged by hybridization to Southern-blotted genomic DNA - data not shown).

Intron 5 contains two different repeated sequences. One is a 4 bp simple repeat that is present in about 200 copies. (Although there is a 400 bp gap in the sequence (see Figure 4.1 and 4.3), as the repeat is present on both sides of the gap, in a nearly perfect tandem array, it is likely that it occupies the gap as well). The repeat is TCT$\frac{A}{G}$, or TATG and apart from a few positions where the repeat is interrupted by a different short sequence it appears to be in a perfect tandem array. Immediately following this repeat, on both sides of the XbaI site at +3454, another internally repeated sequence, with a longer and less obvious repeat unit, is observed (see Figure 4.3 and Table 4.6). The repeat unit is interspersed with non-repeated (but very AT rich) sequences. Units of this repeat are found very close to the intron/exon boundary with exon 6 and also on the other side of exon 6, in intron 6; intron 6 also contains another, different, repeat unit. The 1.3 kb of XbaI fragment (+3336 to +4670) (which comprises the latter part of intron 5, exon 6, intron 6 and part of exon 7) contains a sequence which is repeated ~20 times in the genome (see the Southern blots of genomic DNA - Chapter 3), but whether it is one of the above two repeat units which is present in other genomic locations is unknown. It could be either, or it could be a sequence which is present in only one copy in the 1.3 kb XbaI fragment but repeated elsewhere in the genome. Isolating other genomic clones that hybridize to this fragment and characterizing the basis for the homology would distinguish between these possibilities.

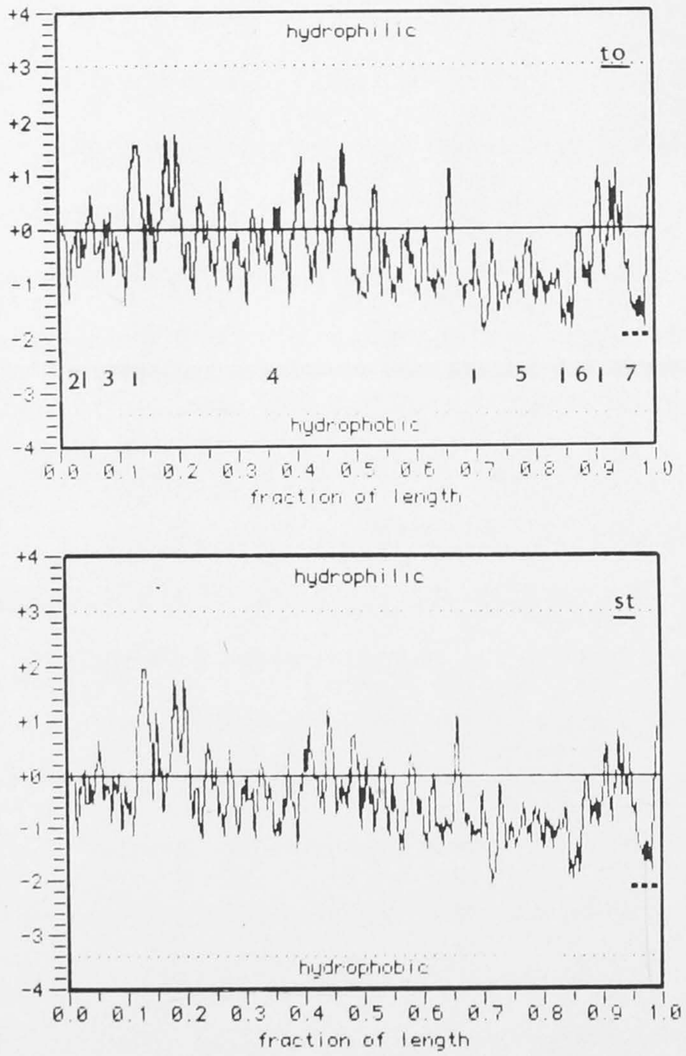
4.2.5 Hydropathy Analysis

The prediction of the primary structures of the topaz and scarlet proteins from their DNA sequence does not by itself provide much information about their conformation or function. Some structural features however can be deduced by hydropathy analysis which indicates whether particular amino acids are located in hydrophobic or hydrophilic regions of the polypeptide; nonpolar residues are preferentially located internally in globular proteins or are associated with membranes, while polar residues tend to be associated with the aqueous environment at the surface of proteins (see Kyte and Doolittle (1982) for a general dicussion on the theoretical basis of this approach). The hydropathy analysis of the topaz and scarlet sequences is shown in Figure 4.7. Windows of 7 or 19 amino acids were moved along the sequence and average hydropathies calculated for each window, with positive values indicating hydrophilicity and negative values hydrophobicity. As expected, the plots for the two genes are extremely similar, further sugesting they perform a similar function in the cell. The topaz protein, as has been noted for the scarlet protein (Tearle, 1986), contains several strongly hydrophobic regions, indicative that the protein may be associated with a membrane. Close to the carboxy terminus, both proteins contain a highly hydrophobic stretch of about 20 amino acids followed by a highly charged region at the carboxy terminus itself. This pattern is a typical membrane-spanning structure and cytoplasmic anchor sequence found in other membrane associated proteins (Rose et al., 1980; Yost et al., 1983).

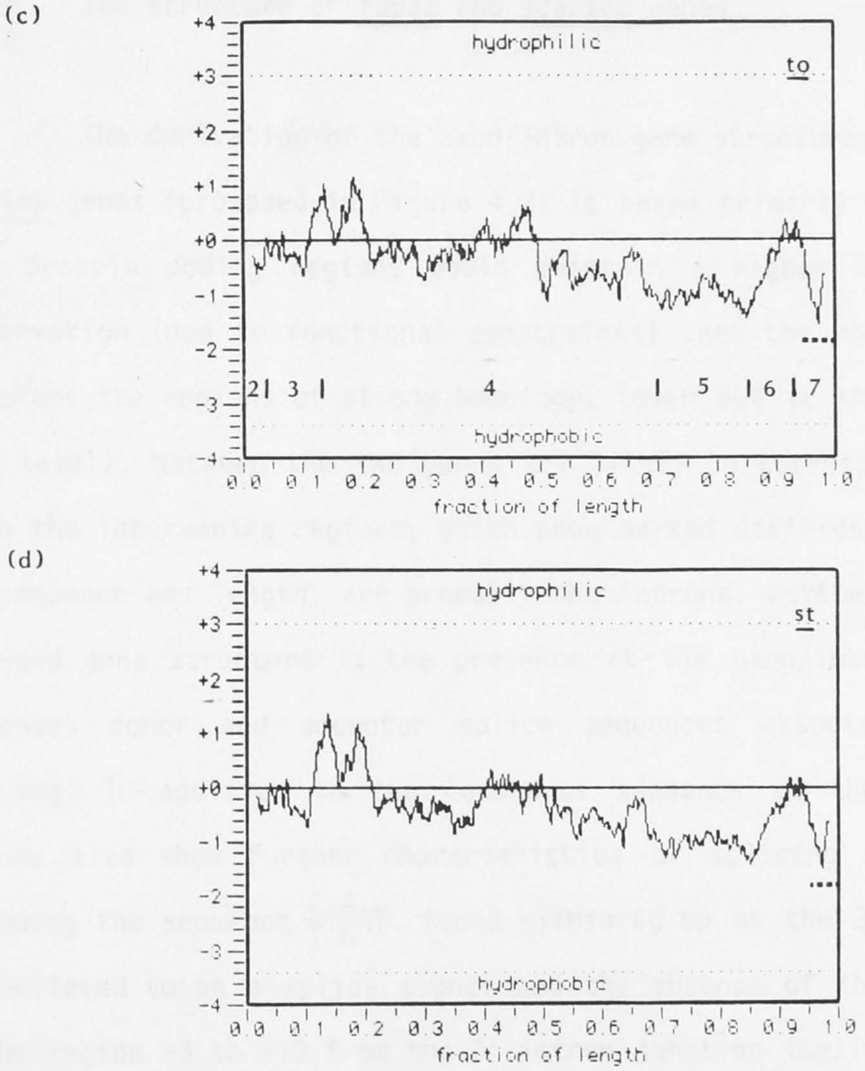
FIGURE 4.7 Hydropathy analysis of the topaz and scarlet genes. Hydropathy values were averaged over windows of 7 residues (a and b) or 19 residues (c and d).

The positions of the exons are marked.

The proposed membrane spanning domains and cytoplasmic anchor at the C-terminus is shown by the broken line.


(a)

(b)


(c)
hydrophilic
to
hydrophobic
2
3
4
5
6
7
+4
+3
+2
+1
+0
-1
-2
-3
-4
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
fraction of length
(d)
hydrophilic
st
hydrophobic
+4
+3
+2
+1
+0
-1
-2
-3
-4
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
fraction of length

4.3 DISCUSSION

4.3.1 The structure of *topaz* and *scarlet* genes

The derivation of the exon/intron gene structures of the *topaz* and *scarlet* genes (proposed in Figure 4.2) is based primarly on the assumption that protein coding regions would maintain a higher level of sequence conservation (due to functional constraints) than the non-coding introns; therefore the regions of strong homology, (over 80% at the predicted amino acid level), between the two genes are likely to correspond to the exons while the intervening regions, which show marked differences in both their DNA sequence and length, are probably the introns. Further support for the proposed gene structure is the presence at the exon/intron boundaries of consensus donor and acceptor splice sequences associated with intron splicing. In addition to the consensus sequence at their 3' ends, the introns also show further characteristics of splicing acceptor regions, including the sequence $\frac{C}{T}T\frac{A}{G}AT$ found within 60 bp of the 3' intron junction and believed to be a splice signal and the absence of the dinucleotide AG in the region -3 to -19 from the 3' intron junction (Keller and Noon, 1985 and see Figure 4.3). A *scarlet* cDNA clone was also isolated and sequenced (Tearle, 1986). This clone, which only carries a short insert, confirms the proposed exon 5, intron 6 and beginning of exon 6 (which defines one end of the cDNA insert).

As outlined earlier, the question concerning the position of the first exon has only been partially resolved; in *scarlet*, Northern analysis indicates that the first exon is positioned in the 0.9 kb PstI/XhoI fragment (-3.2 to -2.3). Analysis of the DNA sequence of this fragment has shown a number of open reading frames that could potentially code for the first exon, the longest of which codes for 70 amino acids. In topaz, the first exon seems likely to be contained within the 2.6 kb SalI/XhoI fragment (-7.8 to -5.2), as it hybridizes weakly to subclone pstX/B0.9. Further restriction enzyme and hybridization analysis of this region in topaz is required to narrow down the region of homology and then DNA sequencing should show which of the open reading frames in the scarlet fragment corresponds to the first exon. Identification of exon 1 is important not only to complete the comparison and analysis of the primary structures of the topaz and scarlet proteins but also to permit comparisons of sequences upstream to exon 1 in the two genes; this could lead to identification of conserved regulatory signals. Such a comparison would be particularly interesting in light of the significant difference in GC composition found between topaz and scarlet.

4.3.2 Codon preference

There is a striking difference in the GC composition and codon preference between exons 2-7 in topaz and scarlet; topaz has only ~30% GC at third base position relative to ~60% in scarlet (Table 4.3). Biases in codon preference have been observed in many genes for which DNA sequence is available. Although there is no clear answer as to why such biases exist, a number of common features are apparent. In E.coli strong correlations exist between the codon preference of genes, their level of expression and the level of tRNA species. In highly expressed genes, the preferred codons represent abundant tRNAs, a measure which is probably involved in energy saving rather than being involved in control of gene expression. This conclusion was reached following experiments showing that in vitro modifications placing strong promotors in front of genes enabled those genes to be highly expressed, even when they contained rare codons (Holm, 1986). A similar correlation between the levels of tRNA species and codon bias has been found in highly expressed genes in yeast, where the likely level of expression of any gene can be predicted from its deduced codon usage (Sharp *et.al*., 1986). In *D.melanogaster*, pronounced codon bias was observed for a number of abundantly expressed genes, however the frequently used codons in *D.melanogaster* are often different from those used in yeast (O'Connell and Rosbash, 1984) indicating that different species might have different pools of abundant tRNA species. In the eight abundantly expressed *D.melanogaster* genes examined by O'Connell and Rosbash (1984), the codons ATA, ACA and CGG were not utilized. In *topaz*, ATA and ACA are preferentially used over their synonyms ATC and ACC, but CGG is not used at all, while in *scarlet*, all three codons are utilized; ATA is slightly over-represented compared with ATC, ACA is under-represented relative to ACC and CGG is highly utilized (Table 4.3). This abundant use of what are regarded as "rare codons" in *D.melanogaster*, would suggest that *scarlet* be placed among lowly expressed genes; such a placement is confirmed by Northern analysis in which the *scarlet* transcript has been found to be less abundant even than the *white* transcript (Tearle, 1986). In the case of *L.cuprina*, it would be interesting to compare the codon usage of abundantly expressed genes with that of genes such as *topaz* or *white*, that are expected to be expressed at a low level. Such a comparison should soon be possible, when clones of the highly expressed *L.cuprina* cuticle protein genes are sequenced (Skelly, 1985; A.J. Howells, pers. comm.).

Recently, a survey of mammalian cDNA sequences revealed a difference in codon usage pattern between genes expressed in liver and in muscle. This has led Newgard *et.al*., (1986) to propose that codon usage, or more specifically %GC at the third codon position, is involved in tissue specificity, and possibly regulates tissue specific expression. An extension of this interpretation would imply that genes that are expressed in the same tissues would have similar codon usages. Although this is true for scarlet and white (Tearle, 1986) it is obviously not the case for topaz and scarlet even though they are expressed in the same tissues, suggesting that such an interpretation is probably not generally valid.

4.3.3 The introns

The putative introns in topaz and scarlet are located at identical positions along the gene, however they differ quite dramatically in length and even more fundamentally, by the presence of repeated sequences in some of the topaz introns. This difference reflects the basic difference in genome organization observed between the two species, i.e. the apparent short periodicity interspersion of repeated sequences in L.cuprina compared with the long periodicity interspersion characteristic of D.melanogaster (see Introduction, Section 1.11).

There is a strong correlation in topaz between the presence of repeated sequences and the length of the intron. The three long introns in topaz, 1, 5 and 6, are the ones that contain repeated sequences (determined by hybridization and DNA sequencing). In the case of introns 5 and 6, which have been nearly fully sequenced, the increased length in the topaz intron relative to the equivalent scarlet intron is primarly due to the presence of repeated DNA. Whether the ancestral gene from which topaz and scarlet have diverged also contained repeated DNA sequences, and the scarlet gene has lost its repeats or whether topaz has gained its repeated DNA at a later stage in the evolution of these genes is unclear. Amplification and reduction in the copy number of repeated sequences have been reported even in closely related *Drosophila* species (Dowsett and Young, 1982), suggesting the latter possibility, of recent gain in *topaz* rather than loss in *scarlet*, is more likely.

There are at least four different repeated sequences found in the introns of the *topaz* gene. Of these, the 4 bp simple repeat present in intron 5 has also been found in a variety of other organisms and is known as the Bkm, or "Garden of Eden" repeat. The Bkm sequence was first isolated as a minor satellite DNA fraction from the female of the Elapid snake, *Bungarus fasciatus* (Singh *et al.*, 1980) and was shown to be absent in males from the same species. (In snakes the females carry the sex determining chromosome which is equivalent to the Y chromosome carried by males in most species). Sequences homologous to the Bkm satellite were also found in genomes of birds, mammals as well as *D.melanogaster* and often in association with the sex determining chromosome. This observation led to the suggestion that the Bkm sequence is intimately involved in sex determination (Singh and Jones, 1982; Jones, 1983). The Bkm satellite is composed of a number of sequences, out of which only one sequence component is conserved in the other species described above; that sequence is the tetranucleotide TCT$\frac{A}{G}$ (Epplen *et al.*, 1982). Regions from *D.melanogaster* genomic clones which carry the 4bp simple repeat in tandemly arrayed blocks have been sequenced and TCTA was found to be the predominant sequence, although some TCTG repeats were present (Singh *et al.*, 1984 and Simpson, 1984). In the *topaz* gene, the repeat appears to be divided into two parts; the first is composed of what appears as a random mixture of the sequence TCTG or TCTA only. Unlike the situation in *D.melanogaster*, in *topaz* the TCTG tetranucleotide predominates. The second part is composed of the TATG perfect repeat, which begins around nucleotide 2500 (Figure 4.2) and continues until the 400 bp gap in the sequence is reached. On the other side of the gap, the TCT$\frac{G}{A}$ repeat is present again. It therefore seems likely that the missing sequences consist of these repeats.

It was noticed that although λ-clones and plasmid subclones containing the Garden of Eden repeats appear stable, M13 subclones carrying the 1.2 BamH1/XbaI fragment were stable only when the insert was in one orientation; when introduced in the opposite orientation (in order to sequence from the XbaI site toward the BamH1 site, see Figure 4.1) the phage deleted the majority of the repeat, preventing completion of the sequencing of the intron.

4.3.4 A possible interpretation of the *topaz* and *scarlet* primary structure

There is a long way between having a deduced amino acid sequence available for a protein and actually predicting its structure or function. The possible ways a protein can fold are so numerous that it is not possible to predict from the primary sequence the correct structure and even if the complete structure is known, there is not enough information available yet correlating the structures with the functions of proteins to assign a specific function.

Biochemical and genetic studies concerned with the ommochrome pathway have lead to some predictions about the function of the *topaz* and *scarlet* genes. They are thought to be involved in the transport and storage of the brown pigment and brown pigment precursors within and between tissues (Summers *et. al.*, 1982 and see Introduction, Section 1.5). Examination of the hydropathy plots derived from the putative primary structures of the *topaz* and *scarlet* genes (Figure 4.7) reveal hydrophobic regions which could correspond to membrane-associated sections of the protein and hydrophilic regions which could be associated with the aqueous environments present on both sides of a membrane. Such features might be expected if the topaz and scarlet proteins are indeed involved in the transport and uptake of xanthommatin and its precursors. A similar hydrophobicity profile has been found for the putative white protein from D.melanogaster (Tearle, 1986). White affects the ommochrome pathway in a similar way to topaz and scarlet and is also thought to be involved in the transport and uptake of pigments and pigments precursors (Summers et al., 1982); white, in contrast to topaz and scarlet, also affects the pteridine pathway.

The question of the possible structure/function of the topaz and scarlet genes and of their comparison with the white genes from L.cuprina and D.melanogaster will be discussed further in Chapter 7.

CHAPTER 5

ISOLATION AND CHARACTERIZATION OF THE MUTANTS TOPAZ^1 AND TOPAZ^2

5.1 INTRODUCTION

Spontaneous (and induced) mutations have attracted much interest because of their contribution to the understanding of gene function and regulation as well as their importance in evolution. Mutations have long been known to be associated with changes at the DNA level, and following the introduction of recombinant DNA techniques, the molecular nature of the defect in many mutants of different genes has been analysed. Among the higher eukaryotes, D.melanogaster, maize and humans offer the most extensive genetic background for the study of mutation, as a wide range of mutants in these species has been genetically characterized and mapped. However the number of spontaneous mutations so far analysed (even in D.melanogaster) is probably not sufficiently high so as to have fully exposed the range of DNA defects which can be involved in the generation of a mutant phenotype.

One consistent pattern which has emerged, (but which could change as more genes are isolated and studied) is that in D.melanogaster about 50% of spontaneous mutations are associated with the insertion of transposable elements (Spradling and Rubin, 1981; Bender et al., 1983; Zachar and Bingham, 1982). In maize, mutations associated with the insertion of mobile elements have also been found (for example Burr and Burr, 1982; Strommer et al., 1982; Sutton et al., 1984), however their relative contribution to the overall level of spontaneous mutations has not yet been determined. In humans, on the other hand, no mutations have been found so far to be caused by insertion of transposable elements or of dispersed repeats such as the Alu sequences (which are now believed to be mobile - see Section 1.11.) Among the mutations examined in humans, the β-thalassaemias are perhaps some of the best studied because their clinical effect enables clear identification. DNA sequencing of 10 independent alleles has revealed that the β-thalassaemia defect can be due to a large deletion which removes part of the β and δ globin genes (Lepore heamoglobin), to small DNA deletions which cause frame shift mutations or simply to single base substitutions which create stop codons in the protein coding region or affect RNA splice sites (Orkin *et al*., 1982a; Orkin *et al*., 1982b). It will probably be some time before the role of transposable DNA sequences in the generation of spontaneous mutations in humans and other organisms can be fully assessed.

Transposable elements and eye colour genes

The nature of spontaneous mutations in four cloned eye colour genes from *D.melanogaster* (*rosy*, *white*, *vermilion* and *scarlet*) has been investigated and each system characterized to different extents. In the case of *rosy*, 4 out of 6 spontaneous mutants examined are associated with insertions of mobile elements (Cote *et al*., 1986). Four spontaneous *vermilion* mutants have been studied using Southern blot analysis. Three of these mutants were found to be associated with the insertion of transposable elements, while one showed no detectable change. It appears that two of the *vermilion* mutants are caused by the insertion of the 412 transposable element into the same location in the gene, suggesting that there might be an insertion "target site" within the *vermilion* gene (Walker *et al*., 1986a, Searles and Voelker, 1986). The *white* locus is particularly suitable for such an analysis as a relatively large number of spontaneous mutations, affecting eye pigmentation to various extents, are available. Zachar and Bingham (1982) have analysed seventeen spontaneous *white* mutant alleles (some of which arose from other unstable *white* mutants). They found that eleven of these mutants are associated with the insertion of non-*white* DNA sequences; most of these sequences were subsequently shown to be transposable elements (O'Hare *et al*., 1984). Among the exceptions were the $\underline{w}^{sp2}$ mutation, that was found to have a small deletion as well as the insertion of non-*white* DNA sequences, $\underline{w}^{Bwx}$, which has a small deletion and four other mutants that were not associated with any change detected at the level of Southern blot analysis. In *D.similans*, 5 out of 7 *white* mutants are associated with DNA insertions which probably involve transposable elements; one mutant is caused by a deletion and one has no detectable change (Inoue and Yamamoto, 1986). Since the overall number of mobile elements is lower in *D.similans* relative to *D.melanogaster*, this result indicates that it is not the absolute number of elements in the genome which determines their contribution to the generation of mutation. Three spontaneous *scarlet* mutants have been examined by detailed Southern blot analysis of their genomic DNA. Two of these mutants have insertions of non-*scarlet* DNA sequences within the gene region; one insertion is in the putative 5' regulatory region and the other insertion in exon 4. Both of these mutants have altered transcription patterns consistent with the insertions being the cause of the mutations. In the third mutant, some changes in its restriction map were detected, however these were thought to be due to polymorphism and not actually associated with the mutation itself (Tearle, 1986).

The relatively high proportion of transposable elements associated with eye colour gene mutations raises the question of whether certain genes contain preferred target sites for the insertion of transposable elements.

The answer is not yet known with certainty, however certain genes have been shown to be "favoured" targets for transposable elements; for example, the singed locus in D.melanogaster has a high frequency of P-element insertions (Rubin et al., 1982). When the chromosomal positions of different mobile elements, after induction of a simultaneous burst of transposition, was examined by in situ hybridization, a preference by various elements for particular chromosomal sites was observed (Gerasimova et al., 1984). This apparent presence of preferred target sites for the insertion of transposable elements should caution against interpretations regarding the role of transposable elements in the generation of spontaneous mutations in organisms where only a few genes have been examined, since those genes might have been unfavourable for insertions and not be truly representative of the genome in general.

Spontaneous mutations in L.cuprina: the topaz mutants

In L.cuprina many spontaneous mutations have been identified and genetically mapped (Maddern et al., 1986). However at the time that this work commenced, nothing was known about the presence of transposable elements in this species or about the molecular nature of the mutations. A major aim of the work was to improve our understanding in these areas.

There are two spontaneous topaz mutants available, called $topaz^1$ and $topaz^2$ both of which were isolated in 1973 (for their affects on the ommochrome pathway see section 1.6). $Topaz^1$ was found as a group of yellow-eyed flies which developed from larvae collected in the field from a struck sheep; $topaz^2$, which has partially pigmented orange eyes, appeared in a laboratory stock which had been established from field flies caught in 1972. Unfortunately more details about their origin are not available; in particular, there is no information concerning the progenitor lines from which the mutants have arisen (G.G. Foster, pers. comm.). The mutant lines available today have been out-crossed over the years with other stocks to avoid reduction in fitness (see section 3.3.3). The mutants alleles were, of course, continually selected on the basis of their phenotype.

Having established the molecular background information concerning the wild-type topaz gene, the system lends itself to study of the two spontaneous mutants, topaz^1 and topaz^2. This chapter describes the experiments designed towards the isolation and initial characterization of these two mutant genes.

5.2 RESULTS

5.2.1 Southern blot analysis of topaz^1 and topaz^2 mutant DNA

The initial step in studying the two topaz mutants was to examine the genomic organization of topaz DNA in the two mutant lines by Southern blot analysis. Subclone ptoHS2.7 (which contains only unique sequence DNA) was used to probe a Southern blot containing HindIII-cut genomic DNA prepared from SWT, topaz^1 and topaz^2 flies (Figure 5.1). In the SWT lane the various polymorphic forms of the topaz gene are detected. In the mutant lanes, hybridization is observed to a 5 kb and an 11 kb fragment for topaz^1 and topaz^2 respectively.

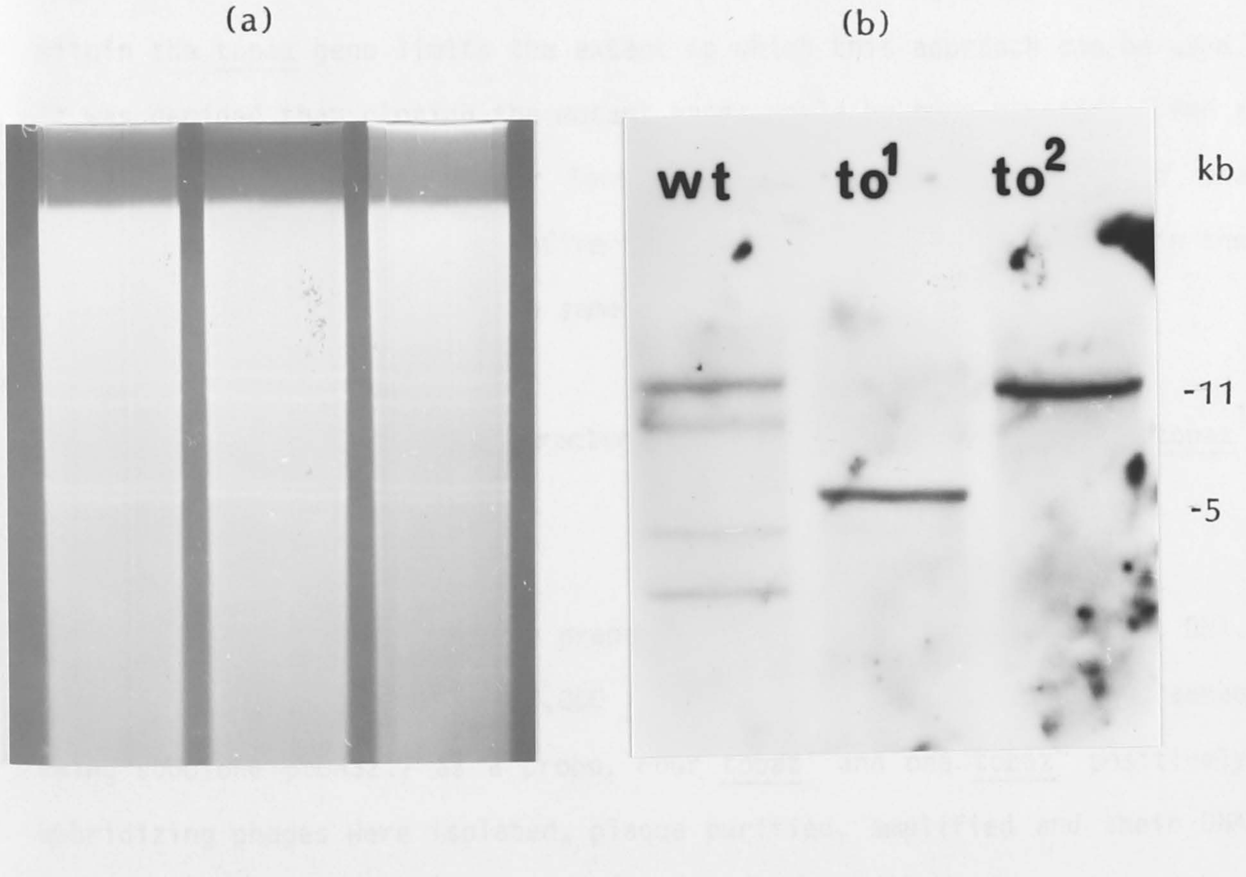
The presence of restriction fragment length polymorphism and restriction site polymorphism in wild-type strains, coupled with the fact that the progenitor strains in which the mutations arose are unavailable, limits the information that can be gathered from such an experiment. Without the progenitor strains, it is difficult to establish whether the

FIGURE 5.1 The genomic organization of the topaz gene region in the mutants topaz^1 and topaz^2.

(a) EtBr-stained gel of genomic DNA from SWT, topaz^1 and topaz^2 digested with the restriction enzyme HindIII.

(b) Autoradiogram of a Southern blot of gel (a), probed with subclone ptoHS2.7.

(Nick translated probe; high stringency hybridization).



changes observed between the mutants themselves, or between the mutants and wild type are due to the presence or absence of a restriction site following a single base pair substitution or whether these changes represent substantial insertions or deletions of DNA sequences. It does however show that the two mutants are different from each other at the DNA level, and that there is no polymorphism in the mutant lines, as these lines are continually selected for the mutant allele through its phenotype. It also shows that whereas the band in topaz^2 DNA is very similar in size to one of the major polymorphic forms in SWT, the band in topaz^1 is not. Although Southern blot analysis using a range of probes covering the whole gene region might have been informative in the study of the genomic organization of the mutant genes, the presence of repeated sequence DNA within the topaz gene limits the extent to which this approach can be used. It was decided that cloning the mutant genes would be more beneficial for a detailed analysis, and further Southern blots using genomic DNA of the mutants were used mainly to confirm that the organization of the DNA in the isolated genomic clones was the same as that of the genomic DNA.

5.2.2 The isolation and characterization of clones carrying the topaz^1 and topaz^2 genes

Genomic libraries were prepared from topaz^1 and topaz^2 adult DNA. Initially, 60,000 topaz^1 and 30,000 topaz^2 recombinant phage were screened using subclone ptoHS2.7 as a probe. Four topaz^1 and one topaz^2 positively hybridizing phages were isolated, plaque purified, amplified and their DNA prepared.

5.2.3 Restriction maps

The topaz^1 clones

The restriction maps of the topaz^1 clones were established as described for the mapping of the wild type topaz gene (Section 3.2.2). An example of a gel containing $\lambda$to^{1}4 DNA digested with various enzymes is shown in Figure 5.2. The fragment sizes obtained show very little similarity to those of the wild type topaz gene. There is also a region in the cloned DNA which appears to be unstable in the phage, a smear of fragments being observed rather than a single band (see, for example, lane 4 in the enlarged photograph Figure 5.2b, in the 4 to 6 kb range). This will be further discussed later. The restriction maps (Figure 5.3b-e) show that all four phage DNAs overlap, covering a region of about 30 kb.

The topaz^2 clones

The restriction map of the topaz^2 clone $\lambda$to^{2}2 is given in Figure 5.3f. By comparison with the topaz$^+$ and topaz^1 maps, it is clear that $\lambda$to^{2}2 does not contain the entire topaz^2 gene but is missing sequences from the 3' region (i.e. beyond co-ordinate 2.1 in topaz$^+$). In order to try to isolate clones carrying the missing region, two additional topaz^2 genomic DNA libraries were constructed in the vector EMBL4. One library used DNA partially digested by Sau3A and the other BamH1-digested DNA. (Southern blots of topaz^2 genomic DNA digested with BamH1 and probed with the 0.6 kb XbaI fragment (4.6 to 5.2 in topaz$^+$), which contains exon 7, showed a fragment of about 14 kb (data not shown), which is an ideal size for directly cloning into the λ vector using size-fractionated DNA). In all,

FIGURE 5.2 Restriction enzyme analysis of the _topaz_^{1} clones. (Example, clone $\lambda$to^{1}4).

(a) EtBr-stained gel. DNA from clone $\lambda$to^{1}4 was digested with the following restriction enzymes S, SalI; S/B, SalI+BamHI; B, BamHI; S/E, SalI+EcoRI; E/B, EcoRI+BamHI; S/H, SalI+HindIII; H, HindIII; H/B, HindIII+BamHI; H/E, HindIII+EcoRI.

Marker - λ/HindIII

(b) An enlarged photograph of part of gel(a) focussing on the region of deleting DNA.

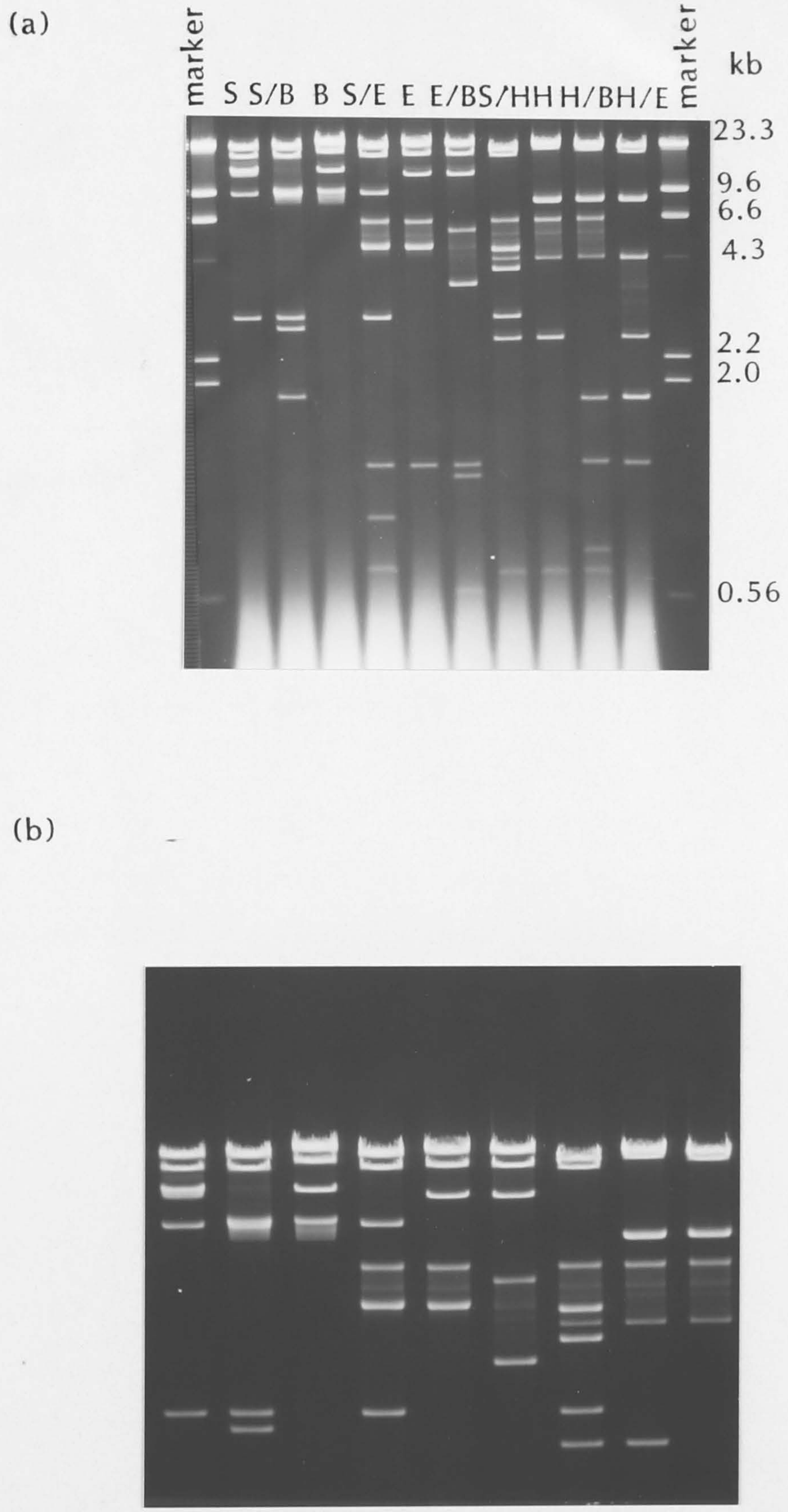

(a)
marker
S
S/B
B
S/E
E
E/B
S/H
H
H/B
H/E
marker
kb
23.3
9.6
6.6
4.3
2.2
2.0
0.56
(b)

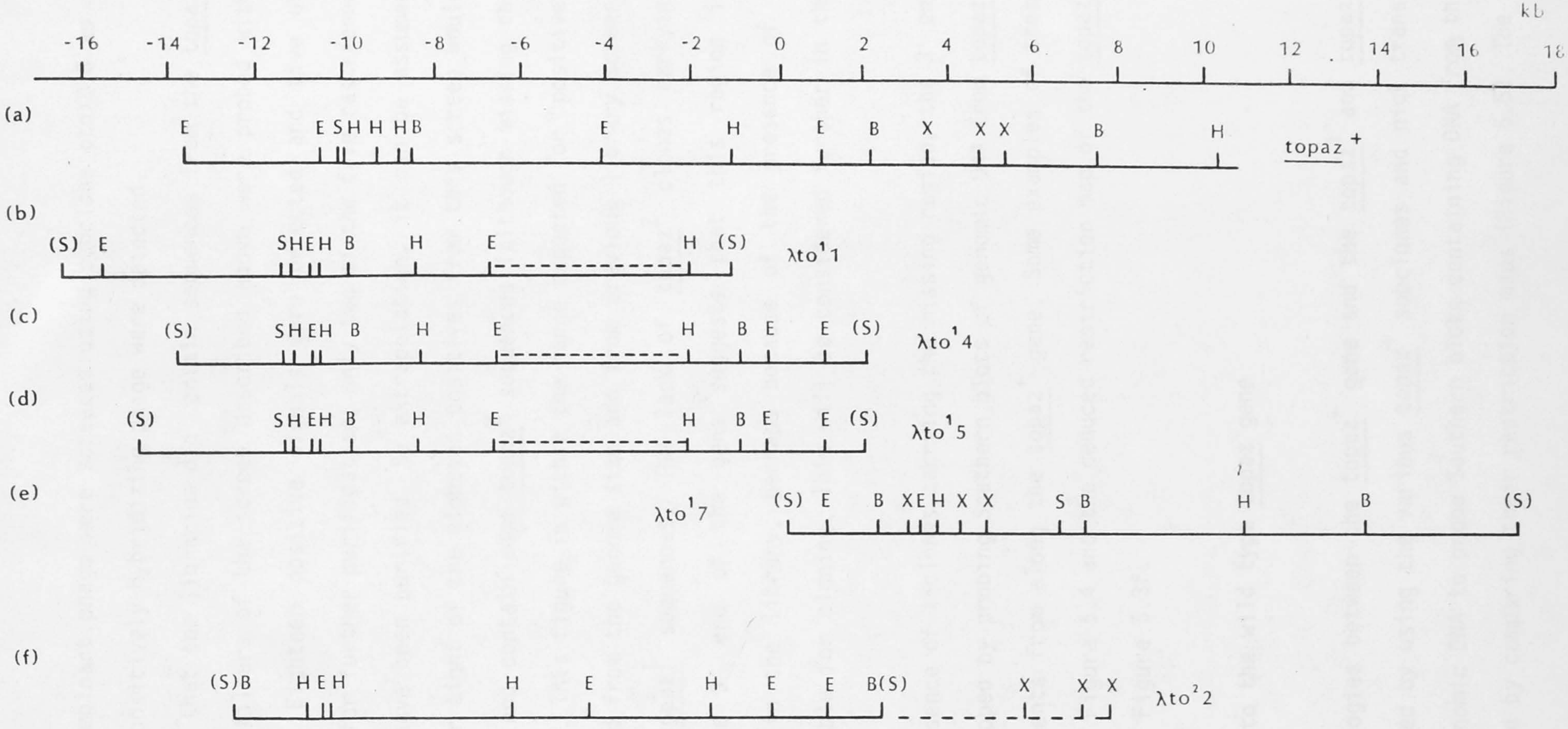
FIGURE 5.3 Restriction maps of the topaz^1 and topaz^2 clones.

All clones are carried in the vector EMBL3A; the SalI site, (S) at the ends of all inserts come from the poly-linker of the vector.

Restriction enzymes used: E, EcoRI; H, HindIII; B, BamHI; S, SalI; X, XbaI (XbaI sites were determined only for the +1.0 to +7.0 region in λto^{1}71).

The broken line in clones λto^{1}1, λto^{1}4 and λto^{1}5 represents the deleting region (see Figure 5.2).

Clone λto^{2}2 extends only up to the (s) site at +2.0. The broken line represents the position of XbaI sites deduced by Southern blot analysis of topaz^2 genomic DNA (see Figure 5.4).


kb
-16
-14
-12
-10
-8
-6
-4
-2
0
2
4
6
8
10
12
14
16
18
(a)
E
E S H H H B
E
H
E
B
X
X X
B
H
topaz +
(b)
(S) E
SHEH B
H
E
H (S)
λto¹1
(c)
(S)
SHEH B
H
E
H
B E
E (S)
λto¹4
(d)
(S)
SHEH B
H
E
H
B E
E (S)
λto¹5
(e)
λto¹7
(S) E
B
XEH X X
S B
H
B
(S)
(f)
(S)B
HEH
H
E
H
E
E
B(S)
X
X X
λto²2

about 250,000 recombinant phage were screened using subclone ptoX0.6 as a probe, however no positively-hybridizing phage were detected.

To ensure that the libraries did contain sequences from the topaz region, duplicate filters of the screen described above were probed with subclone ptoHS2.7. Eighteen positive signals were detected and five of these were picked for further purification; only two of the five rescreened as positives and have been purified. By extrapolation, it can be assumed that about seven or eight of the eighteen positives from this screen would be true positives and contain some topaz^2 sequences (although missing the 3'- end sequences). This figure is within the range expected for "positive" clones as predicted from the genome size and from previous library screens for topaz$^+$ and topaz^1 sequences. The lack of topaz^2 clones carrying sequences from the 3' end of the gene suggests that this region is under-represented in the library, perhaps because of the presence of a sequence unfavourable for cloning. This will be considered further in the Discussion.

In the absence of λ-clones carrying the missing region, the 3' end of the gene was mapped by probing Southern blots of genomic DNA from topaz^2 with various fragments from along the topaz$^+$ gene. Some examples of these blots are given in Figure 5.4 and the deduced restriction map of the topaz^2 region is shown in Figure 5.3f.

5.2.4 Homology to the wild type topaz gene

The homologies between the topaz$^+$ gene and the topaz^1 and topaz^2 clones were examined by using the various topaz$^+$ subclones and nick translated L.cuprina genomic DNA to probe Southern blots containing DNA from the mutant λ-clones and by comparing their restriction maps (Figure 5.5). The

FIGURE 5.4 The genomic organization of fragments from the *topaz*$^+$, *topaz*^{1} and *topaz*^{2} gene regions.

Autoradiograms of Southern blots of SWT, *topaz*^{1} and *topaz*^{2} genomic DNA:

(a) Digested with EcoRI and probed with subclone ptoEB1.2

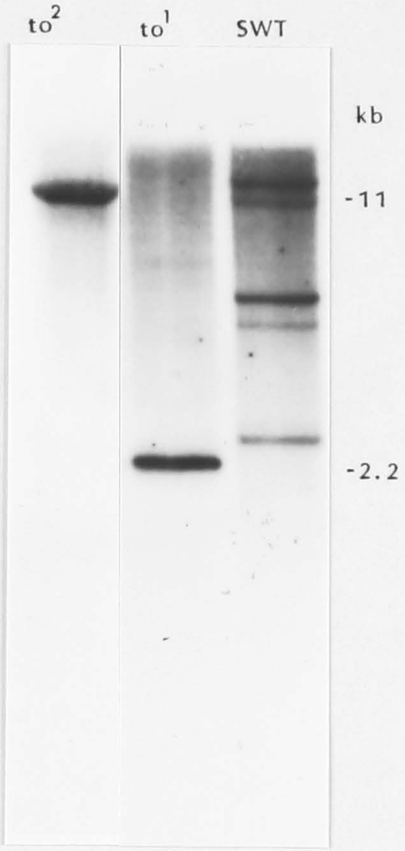
(b) Digested with EcoRI+XbaI and probed with subclone ptoES0.9

(c) Digested with BamHI and probed with subclone ptoES0.9

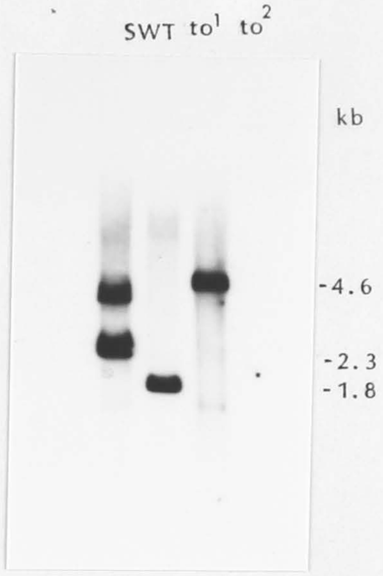
(d) Digested with XbaI and probed with subclone ptoX1.3

(M13 probes; high stringency hybridization)

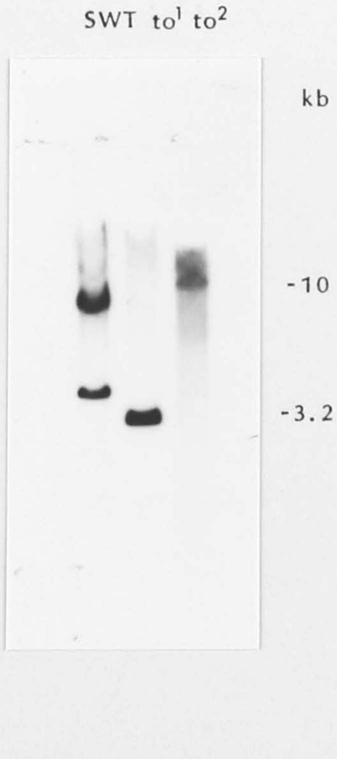
(a)


to2
to1
SWT
kb
-11
-2.2

(b)


SWT
to1
to2
kb
-4.6
-2.3
-1.8

(c)


SWT
to1
to2
kb
-10
-3.2

(d)

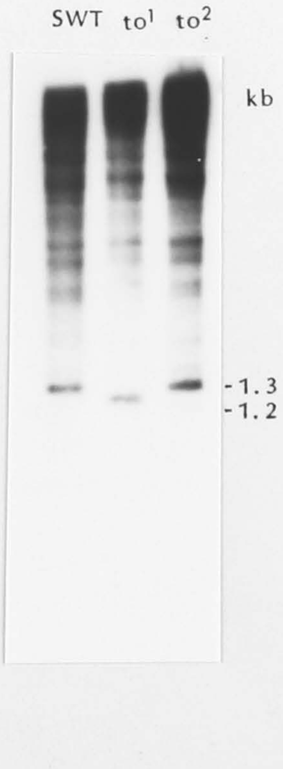

SWT
to1
to2
kb
-1.3
-1.2

FIGURE 5.5 The homology between topaz$^+$, topaz^1 and topaz^2.

(a) An autoradiogram of a Southern blot of the gel shown in Figure 5.2 (clone $\lambda to^1 4$) probed with ptoHS2.7.

(b) An autoradiogram of a Southern blot of the same gel as in (a) probed with nick translated genomic L.cuprina DNA.

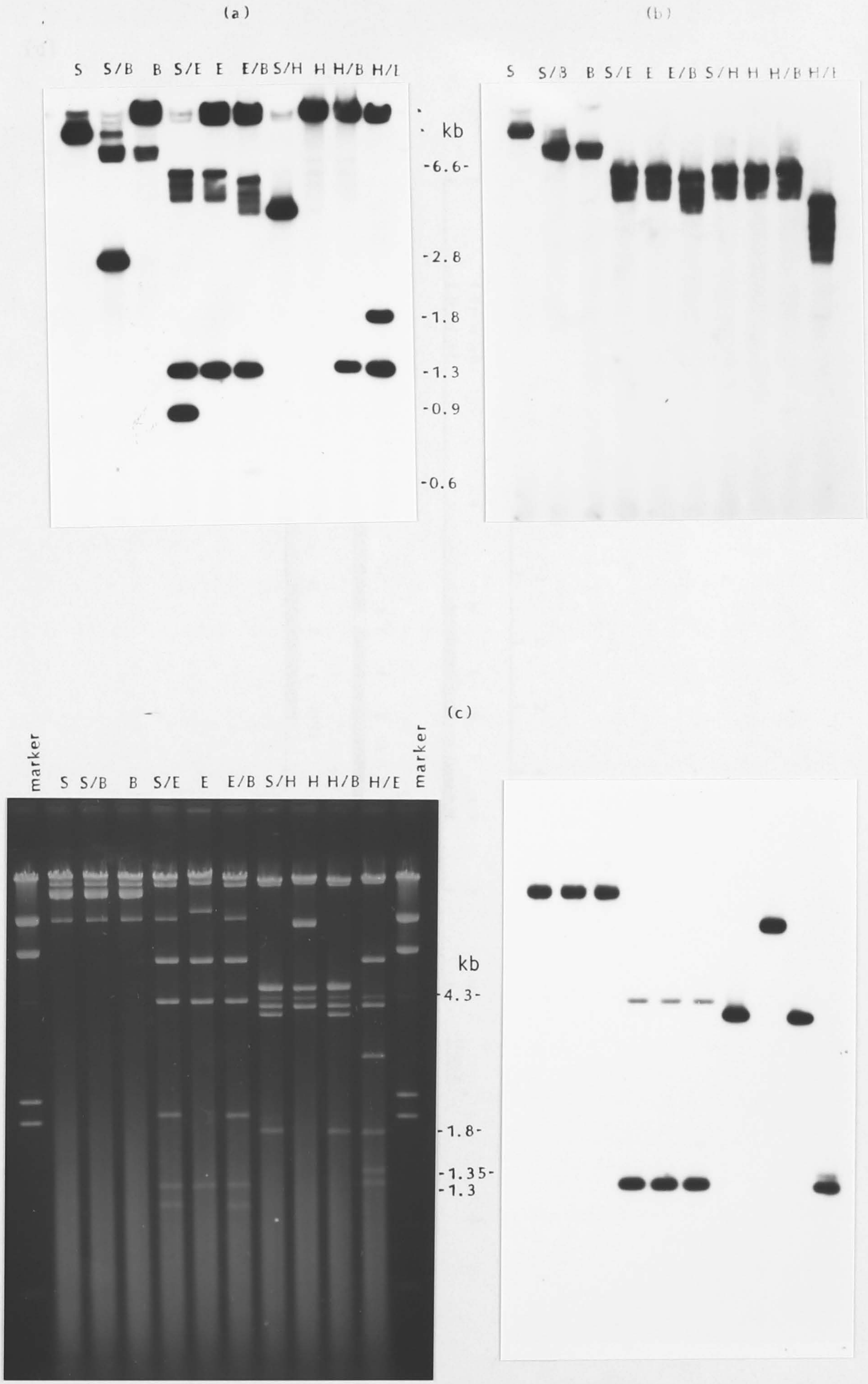
(c) Left: EtBr stained gel of clone $\lambda to^2 2$ digested with the restriction enzymes listed below.
Right: An autoradiogram of the gel on the left probed with ptoHE1.8.

(d) Schematic representation of the regions of shared homology between topaz$^+$, topaz^1 and topaz^2 (next page).

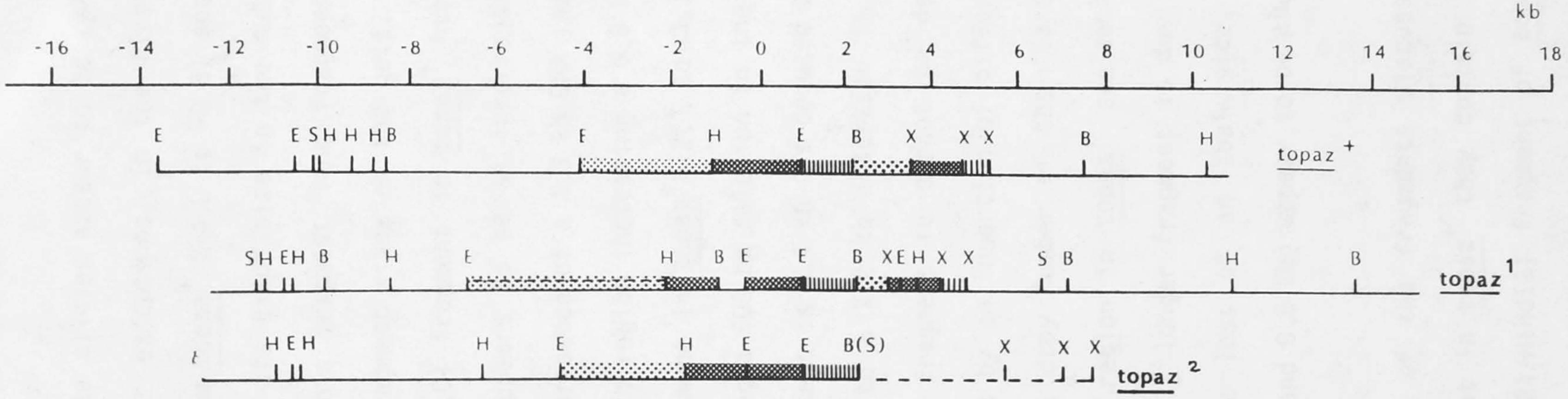
Marker - λ/HindIII.

Restriction enzymes used: S, SalI; E, EcoRI; B, BamHI; H, HindIII; X, XbaI; S/E, SalI+EcoRI; S/B, SalI+BamHI; S/H, SalI+HindIII; E/B, EcoRI+BamHI; E/H, EcoRI+HindIII; H/B, HindIII+BamHI.

(Nick translated probes, high stringency hybridization).


(a)
S S/B B S/E E E/B S/H H H/B H/E
kb
-6.6-
-2.8
-1.8
-1.3
-0.9
-0.6
(b)
S S/B B S/E E E/B S/H H H/B H/E
(c)
marker S S/B B S/E E E/B S/H H H/B H/E marker
kb
-4.3-
-1.8-
-1.35-
-1.3

(d)


kb
-16
-14
-12
-10
-8
-6
-4
-2
0
2
4
6
8
10
12
14
16
18
E
E S H H H B
E
H
E
B
X
X X
B
H
topaz +
S H E H B
H
E
H
B E
E
B X E H X X
S B
H
B
topaz 1
H E H
H
E
H
E
E
B(S)
X
X X
topaz 2

homologies between the three alleles appear to be co-linear along most of the gene, with a number of exceptions. In the first case, the 3.2 kb EcoRl/Hind III fragment from topaz$^+$ (-4.0 to -0.8) appears to be larger in the topaz^1 clone, although its exact size in the mutant clone is hard to establish because it is this fragment (and fragments which contain it) which gave the smear of fragment sizes on the gels, Secondly, the 1.8 kb HindIII/EcoRl (-0.8 to +1.0) fragment in topaz$^+$ hybridizes to a 3.1 kb region in topaz^1. There appears to be an interruption in the region of homology, pointing to the presence of a 1.3 kb DNA insertion. In topaz^2 the hybridization is to a 2.7 kb region indicating a 0.9 kb insertion. Finally, the 1.2 kb BamHl/XbaI fragment in topaz$^+$ (2.1 to 3.3) is only 0.7 kb in length in topaz^1, indicating a 500 bp deletion in the mutant DNA, while in topaz^2 the equivalent fragment is 3.4 kb (as deduced by Southern blotting - see Figure 5.4), pointing to a 2.2 kb insertion in that region. Without having cloned this topaz^2 fragment it cannot be determined whether the change in length is caused by an insertion of different DNA sequences or possibly to an increase in copy number of the "Garden of Eden" sequence (which is located in this region in topaz$^+$ see section 4.2.4). It seems unlikely that the apparently longer fragment is due to a restriction site polymorphism involving the loss of an XbaI site, since the two XbaI fragments downstream (1.3 and 0.6 kb) appear to be the same as in the wild type gene (Figure 5.4).

A common feature of the fragments discussed in the first and second points above is that in topaz$^+$ they contain tandemly repeated DNA sequences. The 3.2 kb EcoRl/HindIII fragment of topaz$^+$ was shown in the previous Chapter to contain at least one type of repeated sequence, organized in a tandem array. The equivalent fragment in topaz^1 is larger, perhaps due to a different copy number of these repeated sequences. As mentioned above, it is this restriction fragment (and others that contain it) that appears to be unstable in the *topaz*^{1} phage, perhaps due to varying deletions of the repeats during propagation. The 1.2 kb BamH1/XbaI fragment of *topaz*$^{+}$ contains a tandemly arrayed block of the "Garden of Eden" sequence (4.2.4). Here again the apparent deletion seen in *topaz*^{1} could be the result of a difference in repeat copy number between the wild type and mutant strains. To ensure that this change in fragment size is genuine and not a cloning artefact (as only one phage covering this region was isolated), subclone ptoES0.9 was used to probe a Southern blot of genomic DNA double digested with the restriction enzymes EcoR1/XbaI (Figure 5.4b). The result clearly shows the predicted 2.3 kb fragment in the SWT lane (plus other polymorphic forms) and the 1.8 kb fragment in *topaz*^{1}.

In addition to these changes in fragment length, there appear to be a number of additional restriction sites in the DNA of the mutants which are absent in the wild type, namely a BamH1 site at position -0.9 (only in *topaz*^{1}), EcoR1 sites at positions -0.3 and 3.7 (the later only in *topaz*^{1}), and a HindIII site at position 4.3 (only in *topaz*^{1}). When the positions of the changes in restriction sites between mutants and *topaz*$^{+}$ are examined with reference to the *topaz*$^{+}$ intron/exon gene structure proposed in the previous Chapter, it appears that all changes occur in intron regions.

5.2.5 Further studies on the putative *topaz*^{1} and *topaz*^{2} insertion sequences

Topaz^{1}

The presence of a DNA insertion associated with the *topaz*^{1} gene suggests the possibility that a transposable element type sequence might be involved. As described in the Introduction (Section 1.12.1), there are a number of features associated with transposable elements; the *topaz*^{1} insertion was investigated for the presence of these features. The region with the insertion contains a BamH1 and an EcoR1 restriction sites, making it convenient to obtain sequences from it in two subclones - pto^{1}HB1.2 and pto^{1}BE0.6 (Figure 5.6)

To examine whether the insertion sequences are repeated elsewhere in the genome a Southern blot of *topaz*^{1} genomic DNA digested with various restriction enzymes was probed with subclones pto^{1}BE0.6, and pto^{1}HB1.2. With subclone pto^{1}BR0.6, about 50 discrete bands of hybridization are obtained (Figure 5.7a) showing that it contains a sequence which is repeated about 50 times in a dispersed fashion in the genome. Thus it shows one of the characteristics of transposable elements. To test for possible transposition as indicated by changes in the size distribution of restriction fragments between stocks, subclone pto^{1}BR0.6 was used to probe a Southern blot of genomic DNA prepared from different *L.cuprina* strains and geographical isolates (Figure 5.7b). Some of the bands appear constant in the different lanes, while others seemed to have changed. This could suggest transposition of the sequence, but could also be a result of restriction site and restriction fragment length polymorphism between the various strains.

Subclone pto^{1}HB1.2 contains a sequence which is repeated at a higher copy number in the genome, judged by its intense hybridization to a wide size distribution of restriction fragments (Figure 5.7c). To estimate the approximate number of copies of this repeat sequence in the genome, subclone pto^{1}HB1.2 was used to probe filters of a wild type *L.cuprina* genomic DNA library. By determining the percentage of positively hybridizing plaques the sequence in subclone pto^{1}HB1.2 was estimated to be

FIGURE 5.6 The subcloning of fragments containing the insertion DNA from *topaz*^{1} and *topaz*^{2}.

(a) The restriction map of clone λto^{1}4; the subclones derived from the region containing the insertion DNA are shown.

(b) The restriction map of clone λto^{2}2; the subclone derived from the region containing the insertion DNA is shown.

Restriction enzymes used: E, EcoRI; B, BamHI; H, HindIII; S, SalI. The SalI sites (S) are derived from the polylinker of the vector EMBL3A.

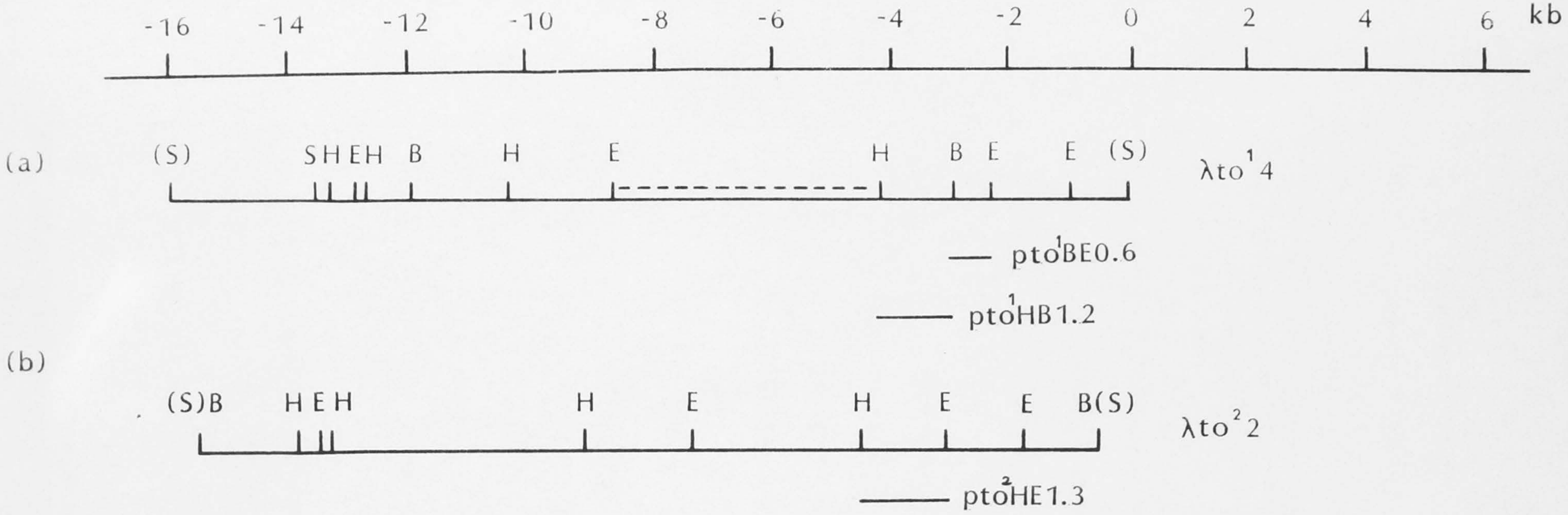

-16
-14
-12
-10
-8
-6
-4
-2
0
2
4
6
kb
(a)
(S)
S H E H
B
H
E
H
B
E
E
(S)
λto¹4
pto¹BE0.6
pto¹HB1.2
(b)
(S)B
H E H
H
E
H
E
E
B(S)
λto²2
pto²HE1.3

FIGURE 5.7 Hybridization of pto^{1}BE0.6 and pto^{1}HB1.2 to L.cuprina genomic DNA.

(a) An autoradiogram of a Southern blot of topaz^{1} genomic DNA digested with the restriction enzymes listed below and probed with subclone pto^{1}BE0.6.

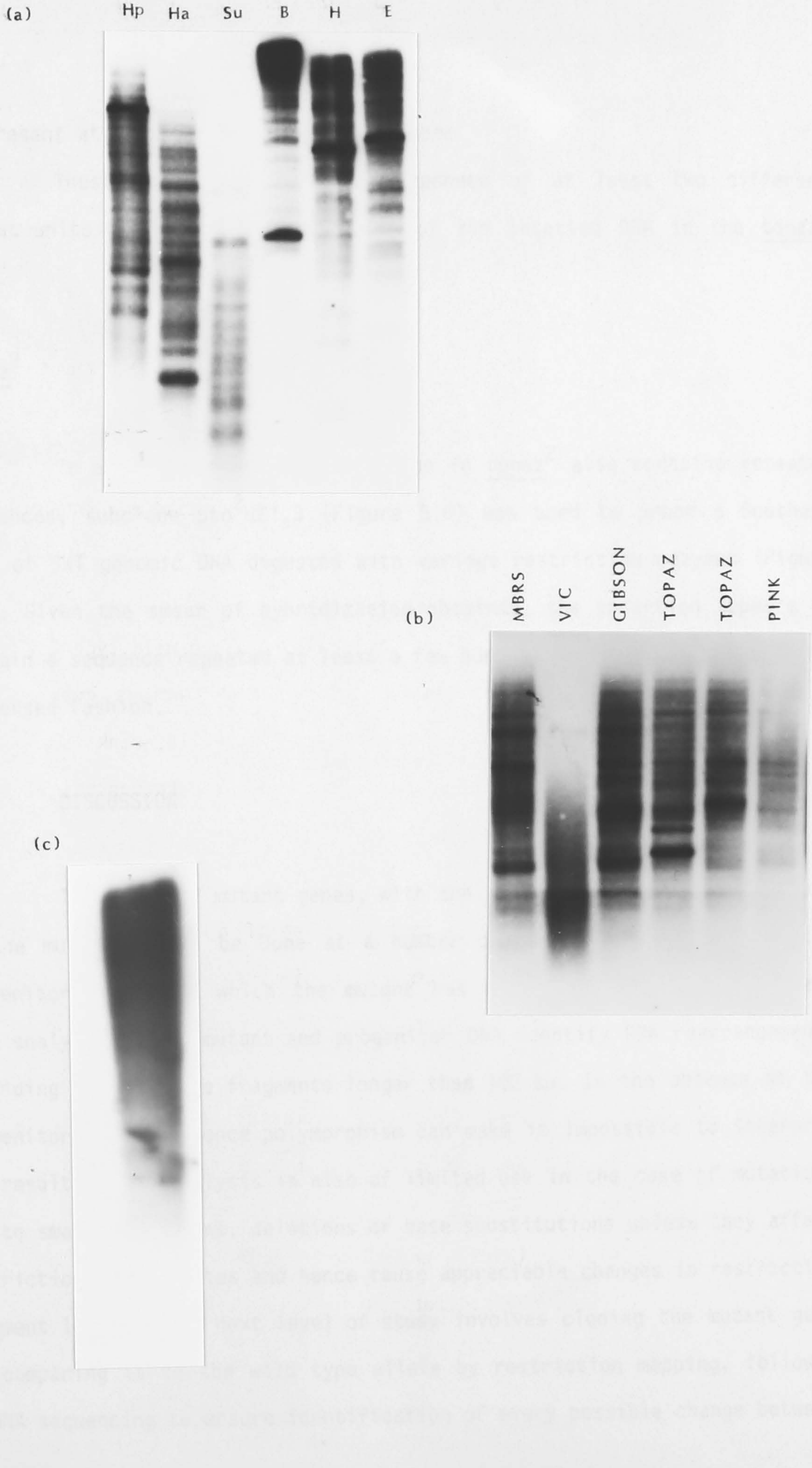
(b) An autoradiogram of a Southern blot of genomic DNA prepared from different L.cuprina strains digested with the restriction enzyme HaeIII and probed with subclone pto^{1}BE0.6.

MBRS, Vic and Gibson are wild type L.cuprina strains isolated from various geographical locations within Australia. Pink is a strain carrying an eye colour mutation resulting in pink eyes.

(c) An autoradiogram of a Southern blot of SWT genomic DNA digested with the restriction enzyme HindIII and probed with pto^{1}BH1.2.

Restriction enzymes used: Hp, HpaII; Ha, HaeIII; Su, Sau3A; B, BamHI; H, HindIII; E, EcoRI.

(Nick translated probes, high stringency hybridization)


(a)
Hp
Ha
Su
B
H
E
(b)
MBRS
VIC
GIBSON
TOPAZ
TOPAZ
PINK
(c)

present at about 4,000 copies per genome.

These results point to the presence of at least two different repeat units within the 1.3 kb region of the inserted DNA in the topaz^1 gene.

Topaz^2

To examine whether the insertion in topaz^2 also contains repeated sequences, subclone pto^{2}HE1.3 (Figure 5.6) was used to probe a Southern blot of SWT genomic DNA digested with various restriction enzymes (Figure 5.8). Given the smear of hybridization obtained, the insertion appears to contain a sequence repeated at least a few hundred times in the genome in a dispersed fashion.

5.3 DISCUSSION

The study of mutant genes, with the aim of identifying the nature of the mutation, can be done at a number of different levels. When the progenitor line from which the mutant has arisen is available, Southern blot analysis of the mutant and progenitor DNA identify DNA rearrangements providing they involve fragments longer than 100 bp. In the absence of the progenitor line, sequence polymorphism can make it impossible to interpret the results. Such analysis is also of limited use in the case of mutations due to small insertions, deletions or base substitutions unless they affect restriction enzyme sites and hence cause appreciable changes in restriction fragment lengths. The next level of study involves cloning the mutant gene and comparing it to the wild type allele by restriction mapping, followed by DNA sequencing to ensure identification of every possible change between

FIGURE 5.8 The hybridization of pto^{2}HE1.3 to *L.cuprina* genomic DNA.

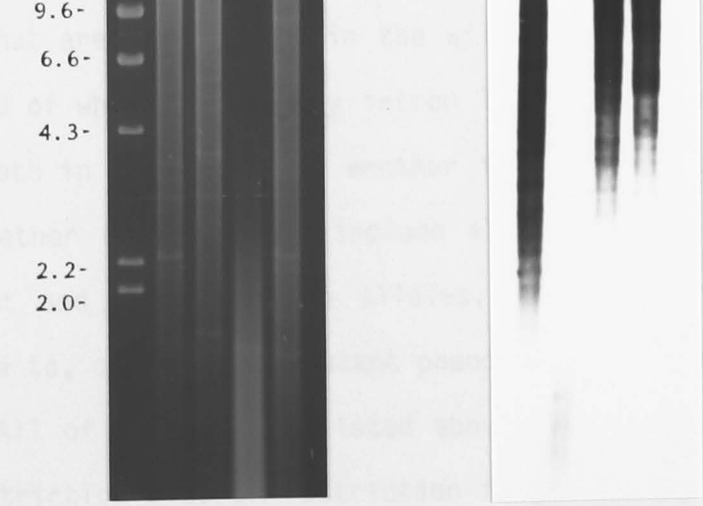
(a) EtBr stained gel of SWT DNA digested with the restriction enzymes: E, EcoRI; H, HindIII; Ha, HaeIII; Hp, HpaII.

(b) An autoradiogram of a Southern blot of gel (a) probed with subclone pto^{2}HE1.3.

Marker - λ/HindIII.

Note that the order of lanes in (b) is in the opposite orientation to that in (a).

(Nick translated probe, high stringency hybridization).



the two alleles. Here again sequence polymorphism can greatly hinder interpretation as to what actually has caused the mutation, and the use of transcript analysis and gene transformation techniques might be necessary in order to provide a definite answer.

5.3.1 The topaz^1 and topaz^2 genes

In this chapter, the preliminary work towards the characterization of the topaz^1 and topaz^2 mutants was described. The genomic organizations of the mutant genes were established and compared to that of the topaz$^+$ allele. In the case of topaz^1, this was achieved by isolating λ-genomic clones spanning the entire gene, restriction mapping, establishing homologies with topaz$^+$ fragments and confirming the proposed organization by Southern blot analysis of mutant genomic DNA. In the case of topaz^2, not all of the gene was cloned, so part of the genomic organization could only be determined by Southern blot analysis using a range of probes from topaz$^+$ and appropriate digestions of genomic DNA from topaz^2.

These comparisons of the three genes have shown that there are a number of changes in the DNA of the mutants relative to the wild type allele. These changes include the presence of restriction sites in the mutants that are not present in the wild type gene, DNA insertions in the distal end of what is probably intron 1 in both mutants, two deletions in topaz^1 (both in intron 5) and another insertion in topaz^2 (in intron 5 as well). Whether these changes include all the changes that exists between the mutant and the wild type alleles, and whether any of these changes contribute to, or cause the mutant phenotype is unknown.

All of the changes listed above appear to be in intron regions, where restriction site or restriction fragment length polymorphisms might be tolerated (unless they affect sequences necessary for accurate intron splicing or gene regulation). Without DNA sequence information it can not be determined whether the appearance of new restriction sites in the mutants is due to single base substitutions or to other DNA rearrangements such as small insertions or deletions. The deletion observed in the *topaz*^{1} gene within the BamH1/XbaI fragment (+2.1 to +3.3) occurs within a region which in *topaz*$^{+}$ is composed almost entirely of a tandemly arrayed simple sequence; this is the same fragment which appears larger in the *topaz*^{2} gene (as judged by Southern blot analysis - see Figure 5.4) suggesting it might represent a fragment length polymorphism brought about by an increase or decrease in the copy number of the simple repeat.

5.3.2 The DNA insertions

The two DNA insertions found in the mutants, (1.3 kb in *topaz*^{1} and 0.8 kb in *topaz*^{2}), have been studied to a somewhat greater extent than the other changes. Both insertions contain sequences which are repeated elsewhere in the *L.cuprina* genome. The insertion in *topaz*^{1} appears to be a composite sequence. Sequences from its right hand side (-0.9 to -0.3) are repeated about 50 times in the genome and show different genomic distribution in different *L.cuprina* geographical isolates, indicating that it might be a transposable element; the left hand side of the insertion (-2.1 to -0.9) contains a sequence repeated thousands of times in the genome. One possible interpretation of these finding is that a double insertion has occurred; there might have been an insertion within an insertion. The insertion in *topaz*^{2} also contains sequences which are repeated in the *L.cuprina* genome at an estimated frequency of a few hundred copies.

The fact that the insertions from the two mutants are located at the same region in the gene is open to several interpretations which are not mutually exclusive: (i) this region could be a target site for DNA insertions and the insertions observed in topaz[1] and topaz[2] are the result of independent events, (ii) the apparent insertions simply represent restriction fragment length polymorphism, are present in other wild type alleles and have no effect on phenotype, and (iii) the insertions are associated with the mutation itself and perhaps topaz[2] is a partial revertant of topaz[1] caused by the partial excision of the topaz[1] insertion, resulting in some restoration of brown pigment production. DNA sequencing of the cloned mutant genes (presented in the following chapter) will address some of the questions raised by the comparison of the wild type and mutant topaz alleles and provide more information about the nature of the insertion sequences.

5.3.3 Regions which have not been analysed

In the cases of both topaz[1] and topaz[2], the analysis of the gene has not included the first exon and the larger part of the first intron. As discussed in the previous Chapter the position of the first exon is still speculative and the results concerning its putative location were obtained too late to be extended to the mutant genes. Therefore work concerned with sequences 5' to the HindIII site at position (-2.1) in topaz[1] or position (-1.7) in topaz[2] was limited. (In fact, attempts to subclone this deleting region in topaz[1] have failed).

As for topaz[2], only part of the gene was successfully cloned, and in spite of constructing and screening three different topaz[2] genomic libraries and over 10 genome equivalents of recombinant phage, exons 6 and

7 could not be recovered; this suggests that there are sequences associated with this region of the gene which are unfavourable for cloning. Under-representation (or perhaps no representation) of certain sequences in L.cuprina genomic libraries appears to be quite common and this is the third example which has been encountered during the work described in this thesis. The first two examples (discussed in Chapter 3) are (i) the lack of representation of certain topaz$^+$ alleles in the SWT library, even though their presence in the same DNA that was used to prepare the library had been shown by Southern blot analysis, and (ii) the absence of sequences homologous to the vermilion gene (presumably the coding regions of yellowish) in the SWT library, while again these sequences were shown to be present in the genomic DNA by Southern blot analysis.

In the case of the topaz gene, the fact that there are sequences associated with this gene which interfere with phage growth became apparent when the poor growth of the topaz$^+$ phage was observed. Additionally, in the topaz^1 clones there is a region which is unstable during phage growth and which deletes part of its sequences; highly repeated DNA sequences are located in this region indicating that they may interfere with the phage replication. Instability of phage and plasmids carrying repeated sequences has been encountered by other workers, for example, during the cloning of D.melanogaster satellite DNA sequences (Lohe and Brutlag, 1986), of the wallaroo satellite sequences (E.S. Dennis, pers. comm.) and of the foldback insertion in the w^c allele of D.melanogaster (Collins and Rubin, 1982); in severe cases they could presumably cause sequences to be unclonable. The region in topaz^2 which could not be cloned is likely to contain a stretch composed of tandem repeats of the "Garden of Eden" sequence. Since these sequences are present in the equivalent position in topaz$^+$ and topaz^1, it is possible that the increase in size of the BamH1/XbaI fragment from this region in topaz^2 could be due simply to an increase in the number of copies of the "Garden of Eden" sequence. Alternatively, it could be due to the presence of another sequence; whatever the explanation, it is likely that it is the presence of these additional sequences which interfere with phage propagation to an extent that causes that region not to be represented in the library. Examination of the restriction maps of the four topaz^1 clones shows that three of them ($\lambda$to^{1}1, $\lambda$to^{1}4 and $\lambda$to^{1}5) end at the same position as $\lambda$to^{2}2, while only one clone spans the 3' end of the gene, indicating that even in topaz^1 this region is somewhat under-represented.

In order to complete the analysis of the changes in the topaz^2 gene, further attempts will need to be made to clone the missing region. Strategies which could be used, include a change in the vector or the host used for cloning. Leach and Stahl (1983) have shown that the stability of phage carrying perfect palindromes can be increased by using a host deficient in the recBC and sbcB products, two exonucleases involved in recombination. Bucheton et_al. (1984) have successfully used this type of host to clone sequences that were unobtainable using other cloning systems. Alternatively, cloning smaller fragments into a plasmid vector might circumvent the sequences actually interfering with the cloning process.

CHAPTER 6

THE ANALYSIS OF THE *TOPAZ*^{1} AND *TOPAZ*^{2} GENES BY DNA SEQUENCING

6.1 INTRODUCTION

The isolation and initial characterization of the *topaz*^{1} and *topaz*^{2} mutant genes has revealed a number of changes both with respect to wild type and between the mutant alleles themselves. The presence of DNA insertions within the gene region of the mutants raises questions about the nature of these sequences, in particular whether they are examples of *L.cuprina* transposable elements. Genetically, *topaz*^{1} and *topaz*^{2} are stable mutants as no revertants have been observed since they were first found, in 1973. If indeed the mutants' phenotype is caused by the insertion of transposable elements, these elements do not have a high rate of transposition, but could perhaps be identified by the presence of structural features associated with transposable elements that have been characterized in other species, i.e. the presence of terminal repeats and the duplication of the integration target site. The insertions are not the only difference between the wild type and the mutant alleles; additional restriction sites are found in the mutants and variation is also apparent in the length of the BamHI/XbaI fragment (which contains the "Garden of Eden" sequence in *topaz*$^{+}$).

The obvious way to investigate the insertions and the other observed changes is to sequence the mutant genes and compare their sequences to the wild type sequence. This Chapter describes the sequence analysis of the *topaz*^{1} and *topaz*^{2} genes.

6.2 RESULTS

DNA fragments from the *topaz*^{1} and *topaz*^{2} gene regions were subcloned into phage M13 vectors and sequenced by the dideoxy chain termination method (see Chapter 2).

6.2.1 *Topaz*^{1}

The sequencing strategy of the *topaz*^{1} gene is shown in Figure 6.1. For reasons discussed in the previous Chapters, fragments 5' to the HindIII site at nucleotide position -840 were not subcloned and have not been sequenced in the mutant genes. The DNA sequence of the *topaz*^{1} gene is (as could be expected from the basic similarity of the restriction maps) quite similar to that of the *topaz*$^+$ gene. Figure 6.2 shows the sequence of the *topaz*$^+$ gene and highlights the changes observed in *topaz*^{1}. These changes include base substitutions, deletions and insertions.

6.2.2 Base substitutions

Single base substitutions are located primarily in introns (for example see Figure 6.3). Three of these substitutions have created restriction sites in *topaz*^{1} which are not present in *topaz*$^+$ (or equally, have abolished a restriction site previously present). A change from T to G at nucleotide position -277 and a change from T to C at nucleotide position 3773 resulted in EcoRI sites and a change from A to T at nucleotide position 4306 resulted in a HindIII site. The fact that these substitutions and therefore these sites are present in the *topaz*^{1} genomic DNA (and are not simply cloning artefacts) was confirmed by Southern blot analysis (see

FIGURE 6.1 The sequencing strategy of the *topaz*^{1} gene.

(a) Restriction map of the *topaz*^{1} gene region.

(b) Sequenced regions of subclones from *topaz*^{1}. Arrows indicate starting points and direction in which sequences were read from the M13 recombinants. Length of arrows indicate the extents to which sequences were read from single reactions.

Restriction enzymes used: E, EcoR1; B, BamHI; H, HindIII; S, SalI; X, XbaI; Su, Sau3A; T, TaqI.

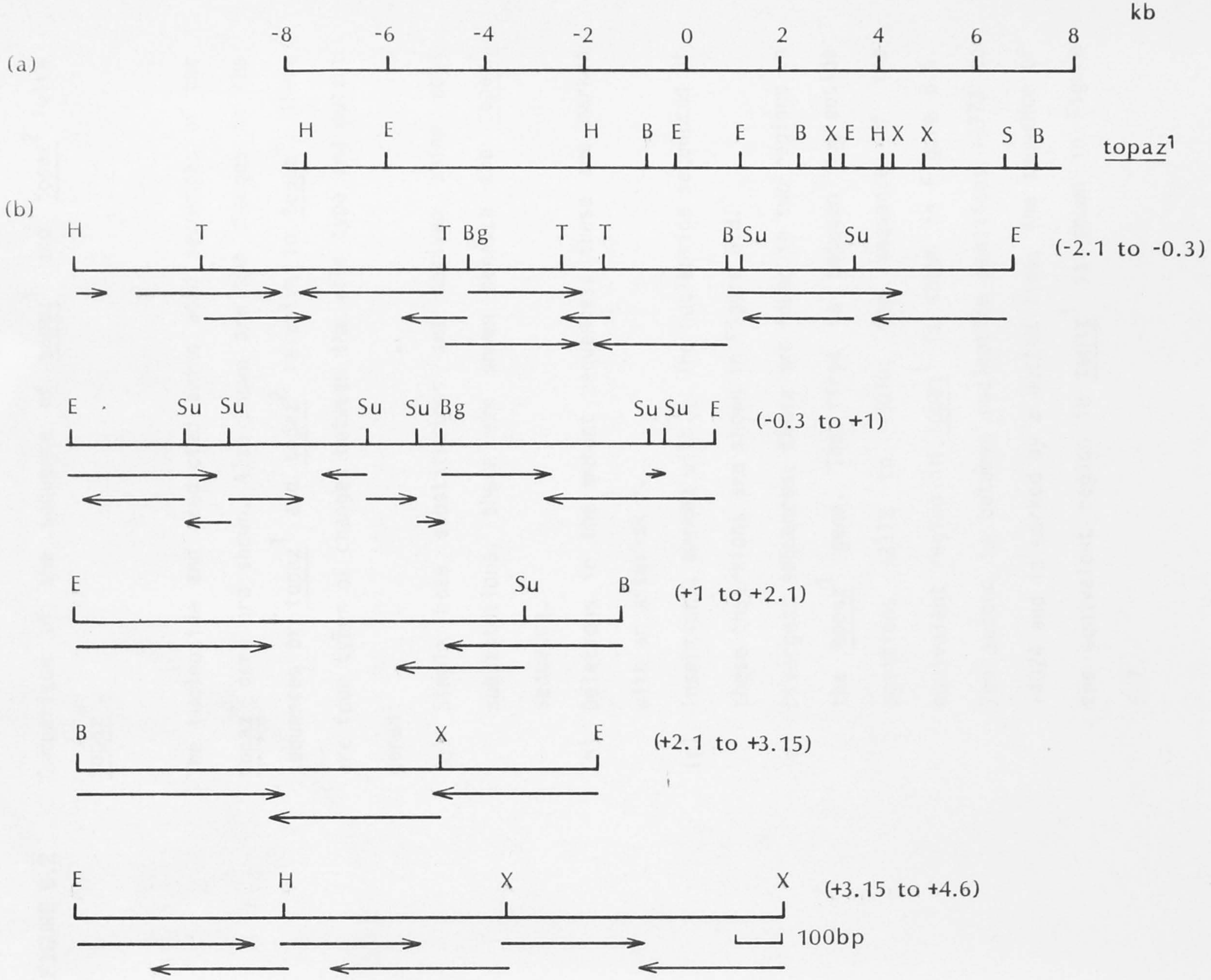

kb
-8 -6 -4 -2 0 2 4 6 8
(a)
H E H B E E B X E H X X S B
topaz1
(b)
H T T Bg T T B Su Su E
(-2.1 to -0.3)
E Su Su Su Su Bg Su Su E
(-0.3 to +1)
E Su B
(+1 to +2.1)
B X E
(+2.1 to +3.15)
E H X X
(+3.15 to +4.6)
100bp

FIGURE 6.2 Comparison of the sequence of *topaz*^{1} and *topaz*^{2} with *topaz*$^{+}$.

The nucleotide and predicted amino acid sequences of the *topaz*$^{+}$ gene are shown. Also shown are the changes in the sequences of *topaz*^{1} and *topaz*^{2} relative to *topaz*$^{+}$. There are four types of changes between the wild type and mutant genes:

(a) Single base substitutions and derived amino acid substitutions; these are shown beneath the *topaz*$^{+}$ sequence.

(b) Deletions in the mutant sequences; these are marked with an asterisk *.

(c) Insertions; marked with ⅄. The nucleotide sequences of these insertions are shown in Figure 6.8.

(d) Divergent sequences; these are found in two regions in the *topaz*^{1} gene. The first is between nucleotide positions +2172 to +3010; the sequence of the equivalent region in *topaz*^{1} is shown in Figure 6.6. The second is between nucleotide positions +3773 to +4024 and is marked by a dotted line; the sequence of the equivalent region in *topaz*^{1} is shown in Figure 6.7.

Where no changes were found in the mutant sequences, only the wild type sequence is shown. For the extent of mutant regions sequenced see Figures 6.1 and 6.13.

Footnote

It should be noted that the coordinates used in this figure are for the *topaz*+ sequence. Since the *topaz*^{1} and *topaz*^{2} sequences contain insertions and deletions their coordinates are different and are shown in Figure 6.1

```
                         HindIII
                         AAGCTTAATTGTGTACTGTTGACATTAAAGTTAAATTTTC

-800 CGAATTTCATTGAAATACAGGTGCTTATGCAGAAAATTAATGAAGAGGAGGAGTTGCAAG

     CGTTAAATAAAGGCCGCTCATTGCGAAATTTTGCAATATCATTTATTGCGATTTAAAACG
                                                       G          to1

     TAGTTGTGCATATTTTTGTTAGGATATTAGTATAATTGACATGATTATGAGCTCTGGGGT
                                               C           A      to1

     TAGGGGTCCGTTTGTGTGGGTGCTAGGTGAAGGAATAAATCGATATCGTCCAAATTCAAT
                                          C        TTG            to1

     AGTATTCGTGCTTTTCACTCAATTATGTTTAAAACTGCGATCTGTAGTTTGATGATTCTG
                                   TC      T                      to1

-500 ACTTCAAAAAATGAAACATTTATCAATACTTCTAAAGTAGGACTAAATGAATTTTTTACC
                                                 C                to1

     TTCTGTGGAAGTGATATTTATTGACTTCCAAAGAGGCAAGATGTTACTCTTTTTGTAGTA

     AAATTTCTAACTAATTTTATTTTATTTAAATTTATTTATGGAAGTTATTTTTTTTTCTGG
                                               1                    2 to1

     GTGTGTCTATTGAAATCATTTCTGGTATTTTCGCAATCTAAATAATTCTTTCAGACATAT
                       A         A    T  A C   GAATTC             to1
                                               GAATTC             to2
                                                EcoRI

     GTATGTAGTTTCGAAAAGAGACCTTTTAAAAATCAGCAAAATCGGTCCATACGACGTTAT
         A                   A A                  ************    to1
        _A         G         A A       T                          to2

-200 GACCAAAAGTTCAAGACAACAACGTATAGAGTTATATGGTAAATTTTCTTGGGCTTAGCT
                                               A                  to1
                   A                                              to2

     TCCTTCTTTATAAAGCAAAACAGAGAATGAATAAAGCATAACATATGAAGCACTACGTTG
      G                                CA             A  G     C  to1
     GG                                                           to2

     TTATCTGATATTTCATCGTTTTAGATTTTCGACGATCATGTTGATCTCTTAAAAACGCAA
         AA T   A  T  T   C                            3      A   to1

     CTTTACTAAAGAAAAAAAGAAAACATTTTAGTGATAAAAATATTTTATTTGATCTATTTT
         T        *G   A      C    A                              to1

     TACAGTGGTTCAGGAAAAACAACATTAATGTCTGTACTTGCATATCGCCAACCAGGTAAT
          aGlySerGlyLysThrThrLeuMetSerValLeuAlaTyrArgGlnProV

+100 TAAAGTTTAGATTTTATTATTTAAATAATCGACAGACGTAGAAACTCGTGAAGACCAAAG
                                                       G   T  to1

     ATTAGATAATTAAAAATAAAAAAATATATTTCAATAATTTCCAGTTGGTACCGTAGTTCA
                                                alGlyThrValValGl
       A      G                                                   to1

     AGGTGATATTCTTATCAATGGCCGACGTATAGGACCATTTATGCATCGCATAAGTGGTTG
     nGlyAspIleLeuIleAsnGlyArgArgIleGlyProPheMetHisArgIleSerGlyCy
```

```
      TGTTTATCAAGATGATTTATTTAATGGATCACTTACCGTGGCAGAACATATGCACTTTAT
      sValTyrGlnAspApsGluPheAsnGlySerLeuTyrValAlaGluHisMetHisPheMe
                                    T                                to1

      GGTAGGAGAACTTCATGTCCTTTTAAATTATTTTTAATAATTTTTACATATTTTAGGCTC
      t                                                       AlaL
                   T                                                 to1

+400  TCTTACGTTTAGATCGCCGCGTTAGCAAGCAGGAACGTAAACTTATAATACAAGATCTTT
      euLeuArgLeuAspArgArgValSerLysGlnGluArgLysLeuIleIleGlnAspLeuP
                               T           A                         to1

      TCGAACGTACAGGTCTATTGGGTGCTTCTAATACACGTATTGGTTCGGGAGATGATGAAA
      heGluArgThrGlyLeuLeuGlyAlaSerAsnThrArgIleGlySerGlyAspAspGluL
       G                                                             to1

      AAGTGTTATCGGGTGGTGAACGTAAACGTTTAGCTTTTGCTGTGGAATTGTTAAATAATC
      ysValLeuSerGlyGlyGluArgLysArgLeuAlaPheAlaValGluLeuLeuAsnAsnP
                                                    T                to1

      CGGTGATATTATTTTGTGATGAACCCACCACTGGTTTGGATTCTTATAGTGCTCAGCAGT
      roValIleLeuPheCysAspGluProThrThrGlyLeuAspSerTyrSerAlaGlnGlnL

      TGGTGCAAACCCTTTACGATTTAGCCAAAAAGGGTACCACTATCTTATGCACCATACACC
      euValGlnThrLeuTyrAspLeuAlaLysLysGlyThrThrIleLeuCysThrIleHisG

+700  AACCGTCTTCACAATTATTTGATATGTTTAATAATGTTCTCTTTTTGTCGGAGGGCAGAG
      lnProSerSerGlnLeuPheAspMetPheAsnAsnValLeuPheLeuSerGluGlyArgV

      TGGCCTTTACTGGTTCACCACAAAATGCTTTGGATTTTTTTGCTCAAAATGGTTATAGAT
      alAlaPheThrGlySerProGlnAsnAlaLeuAspPhePheAlaGlnAsnGlyTyrArgC
                          CAT                                        to2
                          His

      GTCCAGAGGCCTATAATCCGGCCGACTATTTAATAGGTGTACTAGCCTCCGATCCAGGTT
      ysProGluAlaTyrAsnProAlaAspTyrLeuIleGlyValLeuAlaSerAspProGlyT

      ATGAAAAGGCTTCCCAAAGATCAGCTCAATATTTGTGTGATCTATTCGCTGTAAGTTCTG
      yrGluLysAlaSerGlnArgSerAlaGlnTyrLeuCysAspLeuPheAlaValSerSerA
                             ******************                      to1
                       C                                             to2

      CAGCTAAACAGAGAGACATGTTGGTGAATTTGGAAATACATATGGCTGAAAGTGGTGATT
      laAlaLysGlnArgAspMetLeuValAsnLeuGluIleHisMetAlaGluSerGlyAspT
                          EcoRI
+1000 ATCCTTCTGACAAGGAAGTGGAATTCTTTCGTGCTGCTTCTTTGTATTTAAAATTACATG
      yrProSerAspLysGluValGluPhePheArgAlaAlaSerTrpTyrLeuLysLeuHisV
                                                G                    to1
                                                G                    to2

      TTATCTGGTATAGATACACACTGACACTACTGCGTGATCCTAAACTACAGTGGCTGAGAT
      alIleTrpTyrArgTyrThrLeuThrLeuLeuArgAspProLysLeuGlnTrpLeuArgP

      TCTTTCAGAAAATGGCCATGGCCATTATAATAGGTGCCTGTTTTGCCGGTACCACGGTAT
      hePheGlnLysMetAlaMetAlaIleIleIleGlyAlaCysPheAlaGlyThrThrValL
                      T                                              to2
      TGGATCAAATGGGTGTTCAAGCTGTACAGGGTACTCTTTTTGTAATGATTTCTGAAAATA
      euAspGlnMetGlyValGlnAlaValGlnGlyThrLeuPheValMetIleSerGluAsnT

      CTTATCATCCCATGTATTCGGTGTTGAATGTCTTTCCTCAAGGATTTCCATTATTCATGC
      hrTyrHisProMetTyrSerValLeuAsnValPheProGlnGlyPheProLeuPheMetA

+1300 GTGAAACACGTTCTGGCATGTATTCCACAGCTCAGTATTATATTGGCACTGTATTGGCTA
      rgGluThrArgSerGlyMetTyrSerThrAlaGlnTyrTyrIleGlyThrValLeuAlaM

      TGGTAAGATAAGACAAGTAGTAATTAAATGATGATGATGATGACTTGATACATTAATTAA
      et
                        **                                           tol

      AATTCGTATTTGCATGCATGTTCATTGTCTGGTGGCTGGCTGGCTGACTGTTGATTACAT

      GCTATTAATATTTGGTTTTTCCCCCTTCTTTTTATGTTTTCGTTTTGCGCACGTTTATTA

      TCTAGCTGCCGGGCATGATTATAGAACCATTTCTATTTGTTGTCATTTGTTATTTTATCG
           LeuProGlyMetIleIleGluProPheLeuPheValValIleCysTyrPheIleA

+1600 CTGGCTTAAGACCAACATTTTATGCATTTGCTATAACAGCCATAGCTGTTGTACTGGTGA
      laGlyLeuArgProThrPheTyrAlaPheAlaIleThrAlaIleAlaValValLeuValM

      TGAATGTGGCTACAGCATGTGGTTGTTTCTTTTCGACGGCCTTCGATTCGGTACCACTGG
      ETAsnValAlaThrAlaCysGlyCysPhePheSerThrAlaPheAspSerValProLeuA

      CCATGGCATACTTGGTGCCGGTCGATTATATATTCATGATAACATCTGGCATCTTTATTC
      laMetAlaTyrLeuValProValAspTyrIlePheMetIleThrSerGlyIlePheIleG

      AAATCAGGTAAATAAAAAGAAGTATATTTATATGTGACAGTATGTGTTAATGTAGTCTAA
      InIleSe
                                                          C          tol

      GCTAATTGCAAACCCCCTGAAAATATGCTGTGTTTTAGTACAGTTTTCTGTATATTTCAA

+1900 TGTATGTGTGTTTTTGTGTGCTGTAAACTGTTTAATATCATGATCTTTGTGTAAAACAC
                                            A                        tol

      TTAATTTTCCTTGTCATCCTGTCGTTGTCGTTTTTATATTCTTTACCTTTGTAGAAAGTG
                                             C              G        tol

      CATGTTTGAGTTTGTCATTCCGTTGTAATTTTCCCACCTGTCCGTCTGTCTATCATAGAA
         T                              G                            tol
                                                          BamHI
      ATTATGTGTAATTCTTTGATATAATTTCTGATCCTATAAAATATATTTATTTCGGATCCT
                  4               C                                  tol

      TATAGAAAGCGGAGTTGATTGAGCTATGTCCGTCTGTCTGTCTATCTGTCTGTCTGTCTA

+2200 TCTGTCTGTCTGTCTATCTGTCTATCTGTCTGTCTATCTGTCTGTCTATCTGTCTGTCTA

      TCTGTCTGTCTATCTGTCTGTCTATCTGTCTGTCTATCTGTCTGTCTATCTGTCTGTCTA

      ACTGTCTGTCTGTCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTGTCTGTCTATCTG

      TCTGTCTATCTGTCAGCGTGTCTGTCTATCTGTCTGTCCGTCTATCTGTCTGTCTGATTA

      TGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTA
```

```
+2500 TGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTA

      TGTATGTATGTATGTATG  (Gap of c. 400 bp)

      TGTCTATCTGTATGTCTGTCTATCTGTCTATTGGATATTTTGGTACTTTTTCGTTCTTTT

      CTTAAAATTAAAGCTAAAAACATGTAACGTATAAAAAGAACTTTTTGTTCTTTAAAATGT

+3100 ACTTTTTGAGTTTCCATAATATTGAACCTAGAAACATGAAATTAGGCATGTAAAGCCCAG
                                A                                  tol

      ATTATGTAAGAACCTATTAAAGGGGTTCCTTTTGGTACTTTTTTGTTCTTAAAAAGGTAT
             G                  A                                  tol
      TTTTTATTTTTGTTAAACCTAGAAACTGAAATTAAGCATGTAGCCTGCTTAAGGTACGTT
     ⅄    A A                                                      tol
     5
                                                              XbaI
      TTCCTACTCTTGTACTTTTTGAATTTTCTTTAAACTGAGAAACATAAAATTAAACATCTA
           T                                                       tol

      GAGCTAGAAATATGATATTTCCAGATGTACAGCCAAATATAGAACTTTTACTTGTTTAAA

+3400 CACACTATGCTAAAGGGTACTTACGATTCGGAGCAGCCAATTATGTTTTCTTTTTCATTT

      TTGCAAAAAATTCCATTTAATATGTTTTTTTCTCTTATTTTTCAGTACATTGCCTATAGC
                                                   rThrLeuProIleAl

      ATTTTACTGGACACAATTCTTATCATGGATGTTGTATGCAAATGAAGCCATGACTGCTGC
      aPheTyrTrpThrGlnPheLeuSerTrpMetLeuTyrAlaAsnGluAlaMetThrAlaAl

      CCAATGGACTGGTATTACAAATATCAGTAAGTAATTATTTTGTTTTTTTTTTTTGATTGT
      aGlnTrpThrGlyIleThrAsnIleT

      AACACAATTACAGTTTTTCTGTGATTATGTGTATTTGTACAAAGTGGGAATTTACCTTTA

+3700 TGAATACGAGAATTAAACAGCTAAACAGCTAAAGTTTGTTAAATAGATATTAGATTTTAC
                               ********                             tol

      TTATGAAGAATTTAACGGTTATTGTTGAAACATAAGAAA( Gap of c. 70 bp in wt)
             GAATTC(..........sequence different in tol...........  tol
              EcoRI

      TATTAGCCTACACGCAGAAAAAACATGGTTGTGGTAAACATGATATAAGAGCAACATTTTA
      ............................................................. tol

      TTGTTAAATTTATGATTTTAGTTTAAACATATTATAGTTATTATTTTACTCAATCATATT
      ............................................................  tol

      GAAAGTTATAATTGCCATAGTACAATTAACATTAGTTACAATTAACATAGTATAGTTGTA
      ..................................)             G        A   tol

+4050 ATAACCATGACTTTTGTCTTAATATTGTTGAAACAATAATTATATTACGGTGAATCGTAT
                G *    ⅄        A                                  tol
                       6
      TTTAAGTTATACTTACCGTAAATATTCAAATATTTACAGTTATTTAAATATTTTTAAAAT
        A        G     AA A          G  C       C         G          tol
      CCAGTTAAAATGCATATTATTTGTTTTTGATAACAGTACATTGCTAAAATGGACATAGT
        C C      A        A                         C    A           tol

      ATGATTGTAGTAAGCATAACTTTCGACGACGACTGTTAAAACAATAATTATGTTTCGTA
      G      C C   C                   C G                     GGG   tol

      AATCATATTAAAGCTATAATAGTCATAGACATTAAAATCGTTACAGTAACTAAAATATAT
                AAGCTT    TT   C A T C       T         T             tol
                HindIII

+4350 GGTTAAAATCCAAATTAGAACAAATAACAGTTACAATGAACATGTTCGAATTAACATAGA
       A           *  **                                             tol

      ATAGTTGTAGTAACCATGACTTTTGACTTAATGACTTTTGCTTCTAAGCCAATAATAAAG

      AATGTTAATTTGTTACTTAATTTACCAAAAAACTCACTTCTTAGTAGTGATAAGGTTTAA
              A      T                                               tol

      TTCTCCTTTTAAATAATATGACCCTAAAAAGTTAATTATTTTACTGATATTAATGTTTTC
         C                                                           tol

      CTAGAACATGCAATACATTTTATTTATTACAAAAAATTAATTTAATTGTGTGTATTTCGT
                             C                                  C    tol

+4650 TTCTTATTTTCTTCCCTTTTGTCTACAGCTTGCTTCGAAGAAAGTGAAAATTTGCCATGT
                                  hrCysPheGluGluSerGluAsnLeuProCys
           C        T                                                tol
                           XbaI
      TTTCATACGGGTCAAGATGTTCTAGATAAATACAGTTTTAAAGAAACCAATATCTTTCGT
      PheHisThrGlyGlnAspValLeuAspLysTyrSerPheLysGluThrAsnIlePheArg

      AATCTCTTGGCTATTGTTGGTATTTACTTTGGATTTCATATCTTGGGCTATTATTGCTTG
      AsnLeuLeuAlaIleValGlyIleTyrPheGlyPheHisIleLeuGlyTyrTyrCysLeu
                  GTT                                                tol
                  Val

      TGGCGTAGGGCGCGCAAGATATAAGAGAAAAAGTTTCTGCTGCCGAACATTTCTGGTTAT
      TrpArgArgAlaArgLysIleSTP
                                                     G  A            tol

      ACATATTTATAAAATTTTAAGAACAATTAAAGGATAAAGAATTTAAGTCTATGTGTATTT

+4950 GTGTATTAAATAAATTTACATATGTATTTGTTCGTTTTTTTTTTTGGTACTGATTTATAG

      CAAACATTTATAAAAATACAAAAAAAAAAAAATTAACTGAAATAAAATTTTTAACATTAAT

      TTTTATGATATTATTGCGCGCTAAGCAAGGTATGACTATTTGTCTGTCTGTGGCATTGTG

      TCTATTGTGTTTTCAACTTCAAATCAAATAAATTAACAGCCAAAACGTTAGACATTGCAA

      ATGAATAAGAATTTATTTAACACCAAACAAAGAAATTTATTTGCCATCATCAATTTCATG
                XbaI
      CTTCTTTTATCTAGA
```

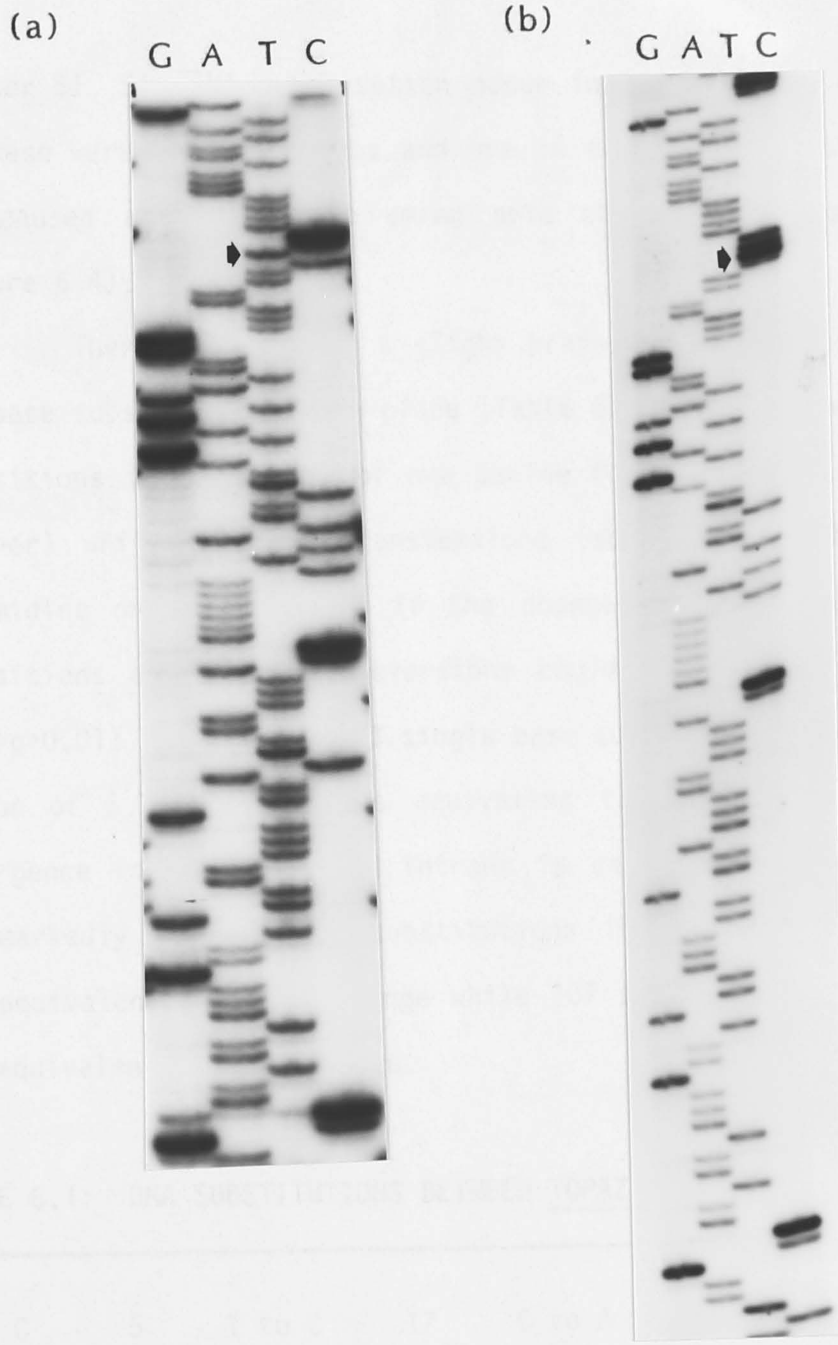
FIGURE 6.3 Single base substitutions in introns (example).

(a) An autoradiogram of a sequencing gel showing part of the nucleotide sequence of intron 6 from topaz$^+$.

(b) An autoradiogram of a sequencing gel showing the nucleotide sequence of the same region of intron 6 as (a) only in topaz^1.

The single base substitution (nucleotide position +4534 in Figure 6.2) is marked with an arrow.

In both gels the nucleotide sequence shown is the complement of the coding strand.


(a)
G A T C
(b)
G A T C

Chapter 5). Six DNA substitution occur in the putative exon regions; five of these were silent changes and one (A to G at nucleotide position 4783) has caused a conservative amino acid change from isoleucine to valine (Figure 6.4).

There seems to be a slight preference in the direction in which the base substitutions take place (Table 6.1); 44% of the substitutions are transitions (substitution of one purine for another or one pyrimidine for another) while 56% are transversions (substitution of a purine for a pyrimidine or vica versa). If the changes were completely random 33.3% transitions and 66.6% transversions could be expected. (χ^2 test shows $0.05>p>0.01$). Altogether 113 single base substitutions were observed over a region of 5.7 kb, which is equivalent to 2% change. If the degree of divergence in the exons and introns is calculated separately the figures are markedly different; 6 substitutions in 1668 nucleotides within exons are equivalent to 0.36% change while 107 substitutions in 4kb of introns are equivalent to 2.7% change.

TABLE 6.1: DNA SUBSTITUTIONS BETWEEN <u>TOPAZ</u>^{1} AND <u>TOPAZ</u>$^+$

A to C	5	T to C	17	C to A	5	G to C	5
A to T	10	T to A	18	C to T	10	G to T	6
A to G	11	T to G	7	C to G	7	G to A	12

In addition to the single base substitutions, there is one region of approximately 240 bp (in intron 6) where the sequence between <u>topaz</u>^{1} and <u>topaz</u>$^+$ completely diverges. These changes are also associated with deletions and will be described in the following section.

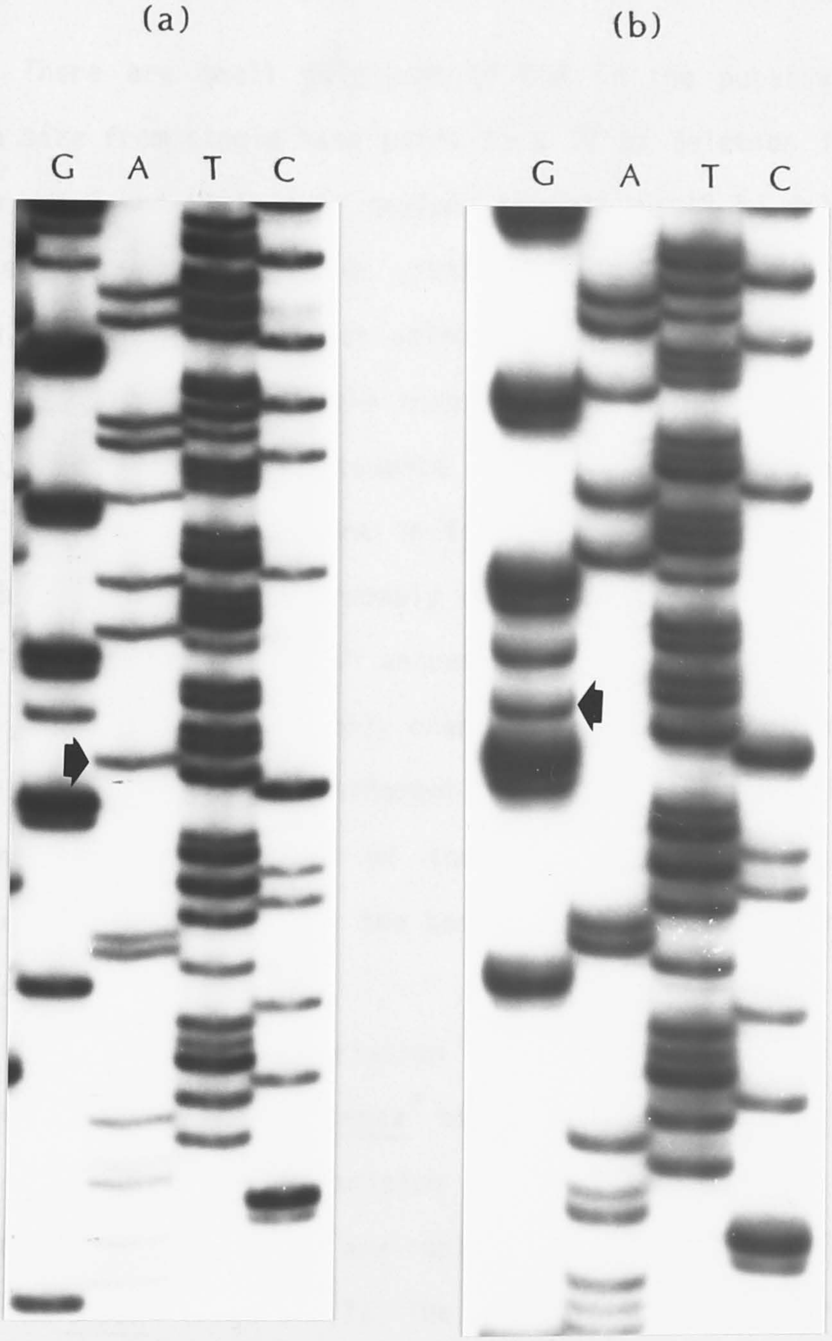
FIGURE 6.4 Single base substitutions in <u>topaz</u>^{1} exons (example).

(a) An autoradiogram of a sequencing gel showing part of the nucleotide sequence of exon 7 from <u>topaz</u>$^+$.

(b) An autoradiogram of a sequencing gel showing the nucleotide sequence of the same region of exon 7 as (a) only in <u>topaz</u>^{1}.

The single base substitution (nucleotide position +4534 in Figure 6.2) is marked with an arrow.

This change caused the amino acid isoleucine to be substituted to valine.


(a)
(b)
G A T C
G A T C

6.2.3 Deletions

There are small deletions of DNA in the putative introns which range in size from single base pairs to a 12 bp deletion in intron 1. One deletion was found in an exon region; this is an 18 bp deletion in exon 4 which removes 6 amino acids without affecting the reading frame. In addition, there are two larger deletions associated with introns 5 and 6, however these are not a simple removal of DNA as they are involved with quite major changes in the sequence around them. In the case of intron 5, the 1.2 kb BamHI/XbaI fragment in topaz$^+$ is only 0.7 kb long in topaz^1. This reduction in size is probably due to a lower copy number of repeat units of the "Garden of Eden" sequence. However, the reduced size of the block of repeats is not the only change involving this region. The internal order of TCTG to TCTA is different in the two alleles (Figures 6.5 and 6.6), and the TATG variant of the repeat is not present in topaz^1. Therefore, strictly speaking, the term deletion is not entirely appropriate in this case.

The second larger deletion is located in intron 6 and results in the 1.3 kb XbaI fragment in topaz$^+$ being only 1.2 kb in topaz^1. Here again the change is not a simple excision of 100 bp but involves the removal of about 240 bp in topaz$^+$ which are replaced by 120 bp of completely different sequence in topaz^1 (Figure 6.7). The presence of these deletions in topaz^1 genomic DNA has been confirmed by Southern blot analysis in which the predicted restriction fragment differences between topaz^1 and topaz$^+$ are observed (see Chapter 5 Figure 5.4).

FIGURE 6.5 The "Garden of Eden" repeat.

(a) and (c)

Autoradiograms of short and long (respectively) sequencing gels showing part of the nucleotide sequence of the "Garden of Eden" repeat from _topaz_$^+$.

(b) and (d)

Autoradiograms of short and long (respectively) sequencing gels showing part of the nucleotide sequence of the "Garden of Eden" repeat from _topaz_^{1}.

Note the different organization of the TCTG and TCTA repeats between the two strains.

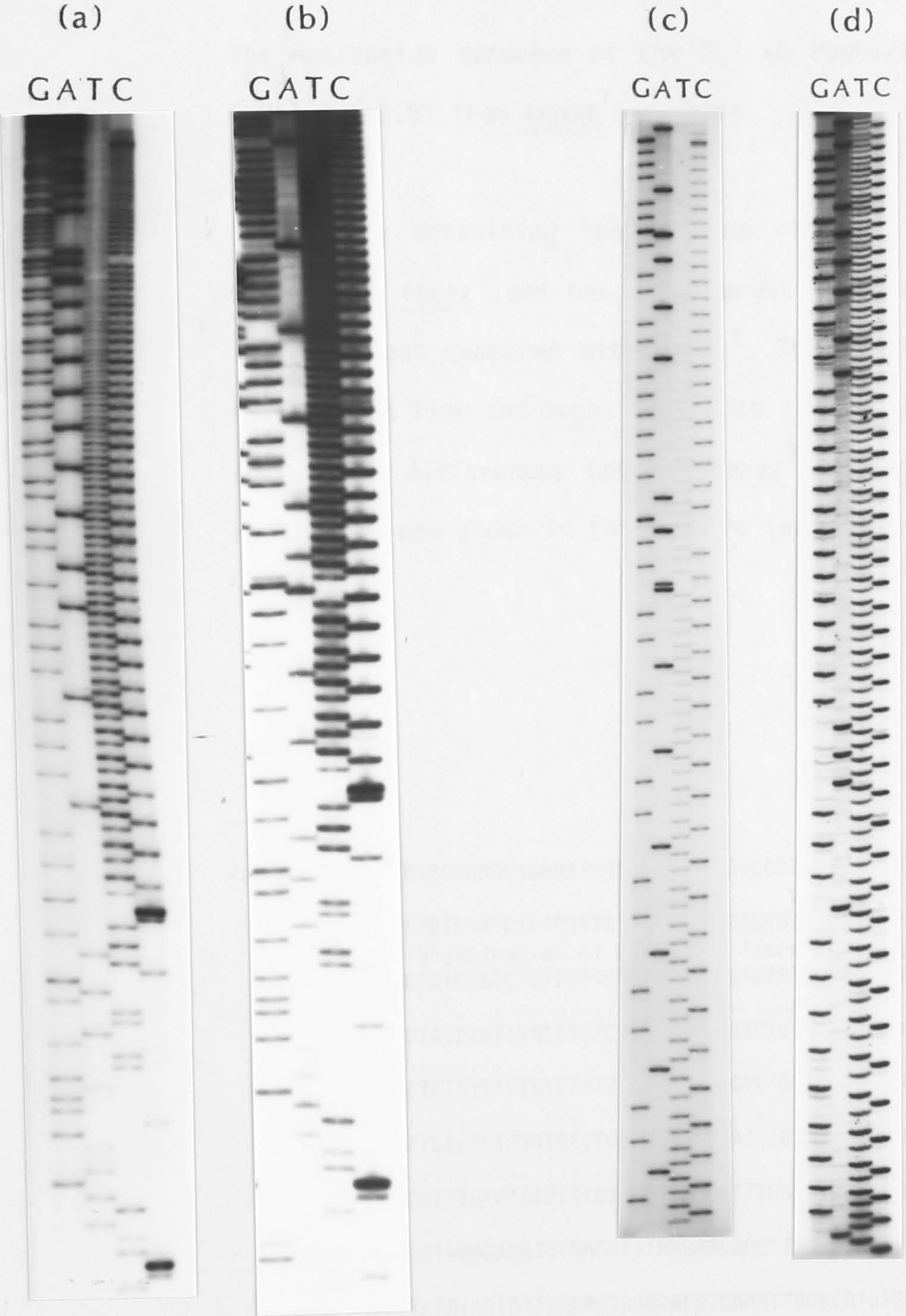

(a)
GATC
(b)
GATC
(c)
GATC
(d)
GATC

FIGURE 6.6 The nucleotide sequence of the 0.7 kb BamHI/XbaI fragment of topaz^{l} containing the "Garden of Eden" sequence.

The nucleotide sequence of the 0.7 kb BamHI/XbaI fragment (+2.1 to +2.8) from topaz^{l} is shown.

The region containing the "Garden of Eden" sequence is shorter in topaz^{l} and has a different organization of the simple repeat compared with topaz^{+}. This region is marked by a dotted line and topaz^{+} sequence can be seen in Figure 6.2. Other differences between topaz^{l} and topaz^{+} in this region are now shown in this Figure (see Figure 6.2).

```
       BamHI
+2133 GGATCCTTATAGAAAGCGGAGTTGATTGAGCTATGTCCGTCTATCTGTCTATCTGTCTGT
                                             (....................
      CAGTCTATCTGTCTATCTGTCTGTCTGTCTATCTGTCTGTCTGTCTGTCTATCTGTCTGT
      .........Organization of repeats different in topaz+........
      CTGTCTGTCTGTCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTGT
      ............................................................
      CTGTCTGTCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTGT
      ............................................................
      CTGTCTGTCTGTCTGTCTGTCTATCTGTCTATCTGTCTGTCTATCTGTCTATCTGTCTAT
      ............................................................
      CTGTCTATCTGTCTGTCTGTCTGTCTATCTGTCTATTTGTCTTTCTGTCTGTCTGTCTAT
      ...........................)
      CTGTCTCTCGTTTATCTGTCTGTCTATTGGATATTTTGGTACTTTTTCGTTCTTTTCTTA

      AAATTAAAGCTAAAAACATGTAACGTATAAAAAGAACTTTTTGTTCTTTAAAATGTACTT

      TTTGAGTTTCCATAATATTGAAACTAGAAACATGAAATTAGGCATGTAAAGCCCAGATTA

      TGTGAGAACCTATTAAAGGGGACCTTTTGGTACTTTTTTGTTCTTAAAAAGGTATTTTTT

      TTAAATTTTGTTAAACCTAGAAACATGAAATTAAGCATGTAGCCTGCTTAAGGTACGTTT
                                                                Xbal
      TCCTTCTCTTGTACTTTTTGAATTTTCTTTAAACTGAGAAACATAAAATTAAACATCTAGA
```

FIGURE 6.7 The nucleotide sequence of the 0.4 kb EcoRI/HindIII fragment of topaz^{l} containing the shorter divergent sequence.

The nucleotide sequence of the 0.4 kb EcoRI/HindIII fragment from topaz^{l} is shown. The region marked by the dotted line is completely different to the equivalent region in topaz^{+} (see Figure 6.2). It is also shorter by about 120 bp in topaz^{l}. Other differences between topaz^{l} and topaz^{+} are not shown in this Figure (see Figure 6.2).

```
          EcoRI(......sequence different in topaz+..................
+3768    GAATTCAGAAATTAGTTTTCATTCATTAGTTGTTCAAACAACTAACTATGTTTCACTAA
         ............................................................
         ATCATATTAATAGTTATAATTACCATAAAAATGTAAACAACCAAAATATGGTTAAAATC
         .......)
         ACTATAATGTTACAATTAACATGGTATAGTTATAATAACCATGGCTTTGTTCTTAATAT

         TATTGAAACAATAATTATATTACGGTGATCGTATTTAAAGTTATAGTTAAAGAAAATAT

         TCAAATGTTCACAGTTACTTAAATATTGTTAAAATCCCGCTAAAATAACTATTATATGT

         TTTTGATAACAGTACATTGCTACAATGAACATAGTGTGATTGCACTAACCATAACTTTC
                                                      HindIII
         GACGACGACCGGTAAAACAATAATTATGTTTCGGGGAATCATATTAAAGCTT
```

6.2.4 Insertions

There are 6 DNA insertions in topaz^1 relative to topaz$^+$ and their nucleotide sequences are shown in Figure 6.8. All the insertions are located within introns. Insertions 3-6 involve only 1-7 bp and have probably been generated by errors during replication, which is likely to be the way the small deletions described in the previous section occured. Insertions 1 and 2 are longer and have been further studied. Insertion 1 is 222 bp long while insertion 2 is 1051 bp long; both insertions are located in intron 1 and are separated by a short DNA sequence of only 22 bp (Figure 6.9).

Hybridization studies presented in Chapter 5 had already indicated that there might be two DNA insertions within the region, (i.e. within the 1.8 kb Hind III/EcoR1 fragment -2.1 to -0.3 in topaz^1). In order to further characterize these insertions, additional λ-clones containing sequences homologous to the two insertions were isolated. Duplicate lifts of 10,000 recombinant phage from the SWT library were screened with subclones pto^{1}HB1.2 and pto^{1}BE0.6. (The BamHI site is located within insertion 2, 520 bp from its 3' end). About 16 phage gave a hybridization signal when probed with subclone pto^{1}BE0.6; it was therefore estimated to contain a sequence repeated about 60 times in the L.cuprina genome. This figure is quite consistent with the number of bands observed on a Southern blot of L.cuprina genomic DNA probed with the same fragment (Chapter 5 Figure 5.7). Subclone pto^{1}HB1.2 was estimated to contain a sequence repeated about 4,000 times in the L.cuprina genome as it hybridized to 10% of the SWT recombinant phage screened. Although this subclone contains both insertion 1 and the left half of insertion 2, the high level of sequence repetition in the genome was initially attributed to a repeat sequence present in

FIGURE 6.8 DNA insertions in _topaz_[1].

There are six insertions in _topaz_[1]; for their positions along the gene see Figure 6.2.

The nucleotide sequence of insertion 1, insertion 2 and the sequence (identical to _topaz_$^+$) which separates them (lower case) is shown on the opposite page.

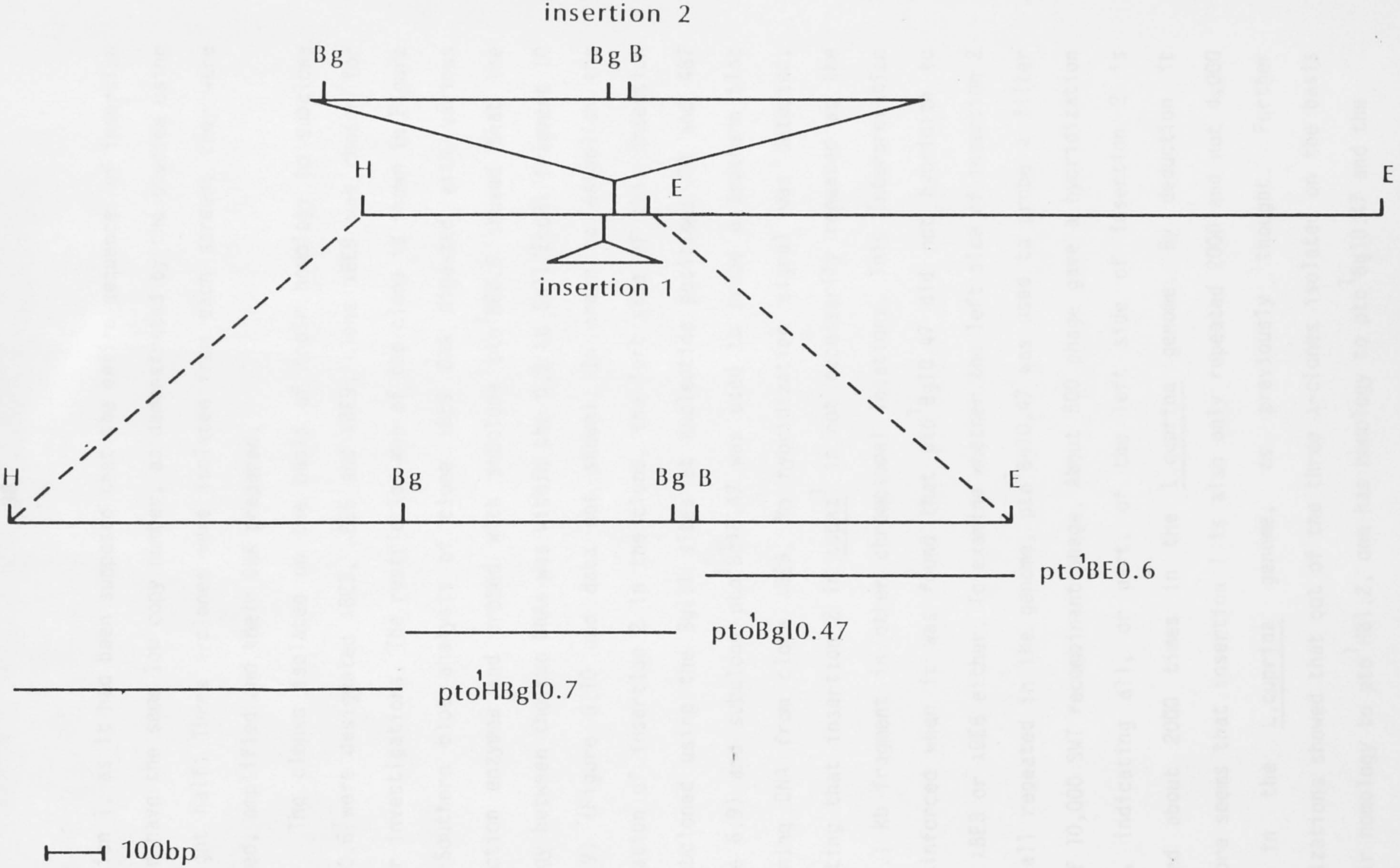
Insertions 3-6 involve between 1 and 7 nucleotides and their sequences are as follows: insertion 3-TCTTATA, insertion 4-T, insertion 5-T, insertion 6-T.

```
   1  TGTAAATATCTTTTAATTCTATAATTATTATACAATTTCACTTCAATTCAATTTCAAAAA

  61  AACAACATAAAAATTACTTCCTGATAATGACTTCAGATATTGTGAATTTACAATCATATA      insertion 1

 121  AAAAGGACATCCCTCTCATGACAAGCTTACGTTAAAATCATCACTTCTTCAGGTCTTTTT
                                          TaqI
 181  CAGTACTAACAAGAAGTAAATTCTACTTCTTTTTCGATTCTT

                 tggaagttatttttttttctgg

                    BglII
 1     CACTTCCAAGATACTTAGATCTAATTTTGACGAAAGTCTGACAGTGGCAAAGAAAGTGCT

       CCAGAGTTTCACCTTACTCTCAATATGCTTTACATTCGTCAAAATGCGCACGTCCAATTT

 120   TGTATAAATGTGCTTGTAATTCTGTGTGTCCACTTAGAAGTAGATACTATAATACTAACC

                                   TaqI                  Sau3A
       TCAGACTTGAACATTTTAAGAAGATTTCTCGACTTGTTCTTATCAGAGTAGGATCTTTGT

 240   GATTCTGTCCACCATTCCAAGCGGCTTTATGGGATTCAGCTTTATTTGTGTCGAATGGTT

       TCCCTTTCGTGCCGACGACTTCTTTTATTTTAGACAAAATTTATGATATTCTTCAACCCT

 360   TCAAAATCTCAACCGTCAGACTACTATGTAGCTGTATGATATGAACTTTTCAATATAAAA

       CATCTATGGTTGTTATTGCTTCTGGGTCAGTAAGTTAGATGAAAATCTTAAACATTAAGT

            BglII                                            BamHI
 480   ATCTCAAGATCTCTTGTAAATTTAAGTGATAGAGTTGATGGTTTCCTTATTGGATCCGTC      insertion 2

            Sau3A         Sau3A,TaqI
       TTGTTGTGATCATTGATTGTTGATCGAACTGATAACCAGAAAGTCACTTATTTAAGTACG

 600   ATATCAAATGGGAGAAATGTATAACATTGGCCAACAAAATTTTAAAAAAAGTAAGACTGT

       TTTATTTGATTATGCCGAATCTTATATTCTCTCCATCAAACCGTATGGCATAAATAGAAT

                                                           Sau3A
 720   GCTCTCCT ATAAAAGCACATTAGTGTCGGTTAAATTAGTTTTCTGACTGATCCTCATAA

       AACTCGGCGGAGAAAAAAGTTTAAAATAATAGAGCGAAAACGAATGTTTTCCGTTTGTTT

 840   TTAAATTTAGGACGCACTATCTCCATAGGCACGGGTACGGACAGTAATTAATCTATTTTA

        Sau3A
       CGATCAGTATCTCTTATAACCTTCACCTTTGTAAGAAGGATATATATGAGTTTGACATTC

 960   CCTCATATTTTCTCAACCCTTCATATTATATAGATTCAGGATTCTTATAGGTGGTGGATT

       GATGTCCGTATGTCTGTCCACCCATTAGTCTG
```

FIGURE 6.9 Restriction map of the 3.1 kb HindIII/EcoRI region (-2.1 to +1) in topaz[1].

The positions of insertion 1, insertion 2 and the subclones derived from them are marked.

Restriction enzymes used are H, HindIII; E, EcoRI; B, BamHI; Bg, BglII.


insertion 2
Bg
Bg B
H
E
E
insertion 1
H
Bg
Bg B
E
pto^{1}BE0.6
pto^{1}Bgl0.47
pto^{1}HBgl0.7
100bp

insertion 1, as it had been supposed that the entire sequence of insertion 2 would have the same low copy number, as demonstrated by the studies using its right half. Three λ-clones were isolated from each screen; they were purified, amplified and their DNA prepared.

The clones isolated on the basis of their homology to subclone pto^{1}BE0.6 were designated λBE3, λBE6 and λBE9; clone λBE9 was chosen for further investigations. The restriction map of the clone is shown in Figure 6.10. Southern blot analysis of clone λBE9 DNA digested with various restriction enzymes and probed with subclone pto^{1}BE0.6 showed that the homology between the two DNAs was within the 0.9 kb SalI/EcoR1 fragment (0 to +0.9) (Figure 6.10 and data not shown). In order to establish the orientation of insertion 2 in the clone, the left side of this insertion was subcloned using the BglII sites at nucleotide positions 18 and 487 (Figure 6.9) and subclone pto^{1}Bg10.47 was used to probe a Southern blot containing DNA from clone λBE9. No hybridization signal was detected, indicating that insertion 2 in *topaz*^{1} is not necessarily repeated as the same 1.1 kb fragment in other chromosomal locations. This interpretation was reinforced when it was found that pto^{1}Bg10.47 did not hybridize to clones λBE3 or λBE6 either. To examine whether the left side of insertion 2 is at all repeated in the genome, pto^{1}Bg10.47 was used to probe a filter lift of 10,000 SWT recombinant phage. About 500 phage gave a hybridization signal, indicating all, or part, of the left side of insertion 2 is repeated about 2000 times in the *L.cuprina* genome. By deduction it therefore seems that insertion 1 is also only repeated 2000 and not 4,000 times in the *L.cuprina* genome, as previously thought. Further hybridizations showed that out of the three λ-clones isolated on the basis of their homology to pto^{1}HB1.2, one has homology to pto^{1}Bg10.47 and the

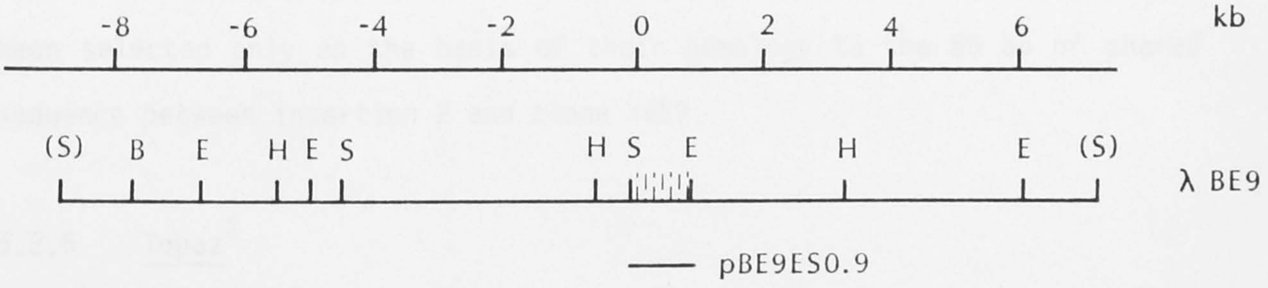
FIGURE 6.10 Restriction map of clone λBE9.

Restriction enzymes used: E, EcoRI; B, BamHI; H, HindIII; S, SalI.

The SalI sites (S) are derived from the EMBL3A polylinker.

The region of homology to insertion 2 from <u>topaz</u>[1] is marked by ¦¦¦¦

The subclone derived from this phage is marked.



other two have homology to subclone pto^{1}HBg10.7, which contains insertion 1 only.

The question of what part of insertion 2 is included in clone λBE9 was addressed by subcloning and partially sequencing the 0.9 kb SalI/EcoRI fragment (0 to +0.9). The sequencing strategy is shown in Figure 6.11 and the nucleotide sequence is shown in Figure 6.12. The sequence homology between insertion 2 and the 0.9 kb SalI/EcoRI fragment from clone λBE9 is restricted to 65 bp (nucleotide positions 668 to 733 and 4 to 69 respectively), out of which 86% are homologous. There is a 250 bp gap in the sequence data; while the possibility cannot be ruled out that further homology in that region exists, this seems unlikely as there is no homology between sequences in insertion 2 and sequences flanking the gap in pBE9ES0.9. A Southern blot containing DNA from clones λBE3 and λBE6 was probed with pBE9ES0.9 to examine whether these clones might also contain the 65 bp shared between insertion 2 and clone λBE9; the result (data not shown) showed hybridization to both clones; since neither of the λBE clones showed homology to the left side of insertion 2, they in fact might have been selected only on the basis of their homology to the 65 bp of shared sequence between insertion 2 and clone λBE9.

6.2.5 *Topaz*^{2}

The *topaz*^{2} gene was only partially sequenced, with the main aim being to examine the regions that were found to be different between *topaz*$^+$ and *topaz*^{1}; specifically the region containing the deletion in exon 4 and the insertions in intron 1. The sequencing strategy is shown in Figure 6.13. The changes between *topaz*$^+$ and *topaz*^{2} are shown in Figure 6.2; single base substitutions are found in both introns and exons. There are 4 base

FIGURE 6.11 The sequencing strategy of subclone pBE9ES0.9.

Sequenced regions from subclone pBE9ES0.9. Arrows indicate starting points and direction in which sequences were read from the M13 recombinants. Length of arrows indicate the extents to which sequences were read from single reactions.

Restriction enzymes used: E, EcoRI; S, SalI; T, TaqI.

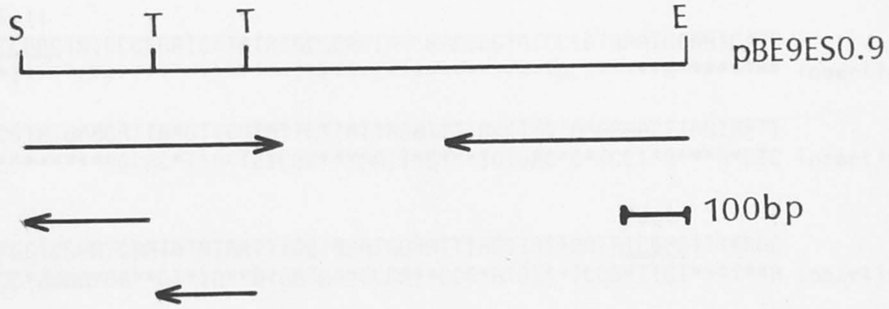


FIGURE 6.12 Comparison of sequences from the 0.9 kb SalI/EcoRI fragment of clone λBE9 with sequences from insertion 2 of *topaz*^{1}.

The partial nucleotide sequence of the 0.9 kb SalI/EcoRI fragment (0 to +0.9) is shown. There is a 65 bp region of homology between the above sequence and the sequence of insertion 2. This region begins at nucleotide position 668 in insertion 2 (see Figure 6.8) and identical nucleotides are marked by an asterisk *. There is no significant homology beyond nucleotide position 733 in insertion 2.

```
     SalI
  1 GTCGACTATGCCGAATCTTATATACCCACTATCAAACCGTATGCTGTAAATGGAATGCTC
665 T*T**T*****************T*T*T*C*************GCA*****A******** insertion 2

    TCCTATAAACATTAAGTTGATATTATTATTAGATTTTAACTTGTAAGAAACTTAATAATT
725 *********AGCAC*T*A*TGTCGG***AATT*G***TCTGAC*G*TCCT*A***A*CTC insertion 2

                                                     TaqI
120 ATGGTCCGATCAATATATAATTTGGTAAATAGAATTTACTTATTAATATCGAATTTAAAC
785 GGC*GAGA*AA**GT*TA**A*AATAG*GCGA**ACG*A*GT**TCCG*TTGT***T**A insertion 2

    TCTTATATTTAAGACAGTAAAGCCATTTTTTTGGAGTGGACCATATATGAGGGGTAAGGT

180 AAATTATTTACCGACCCTCATAAAATTTGGTAGAGAGAATTTGAATCACACAAGACTTAT

        TaqI
    TTATGTCGATTTTATCGCTATACTACCACTTGTATGGGGGCTATACGAAAACGTGGACCG

240 TT (Gap of c. 250 bp)

    GACCCTATGGAAAATTACTTTAATGCTCATAACTGTCTGAAAGAAACATCTATAGTGGTG

550 AAATTCAACAGGAAGATATTTTATAGGAAAATTTTATTGCAGTATACATAGCAATAAAAT

                                        TaqI
    TTTTCGTTAAAAAAGTTTTATAGGAGCTAAAATCATTCGACATAATCAAACCTGTCTACT

710 AACTTTTATAAGTGTCAGTTGAAAAATGACCCTACCCCATATAAAGGCGTCATGAAATAA

    CTCCCCATAACAAAAATTAAATAAGTCCTCAAATTTGGGAATATTTTTCATTATGCCGAT

830 TATATAATACCCCACTAAACAATTTTTGCC
```

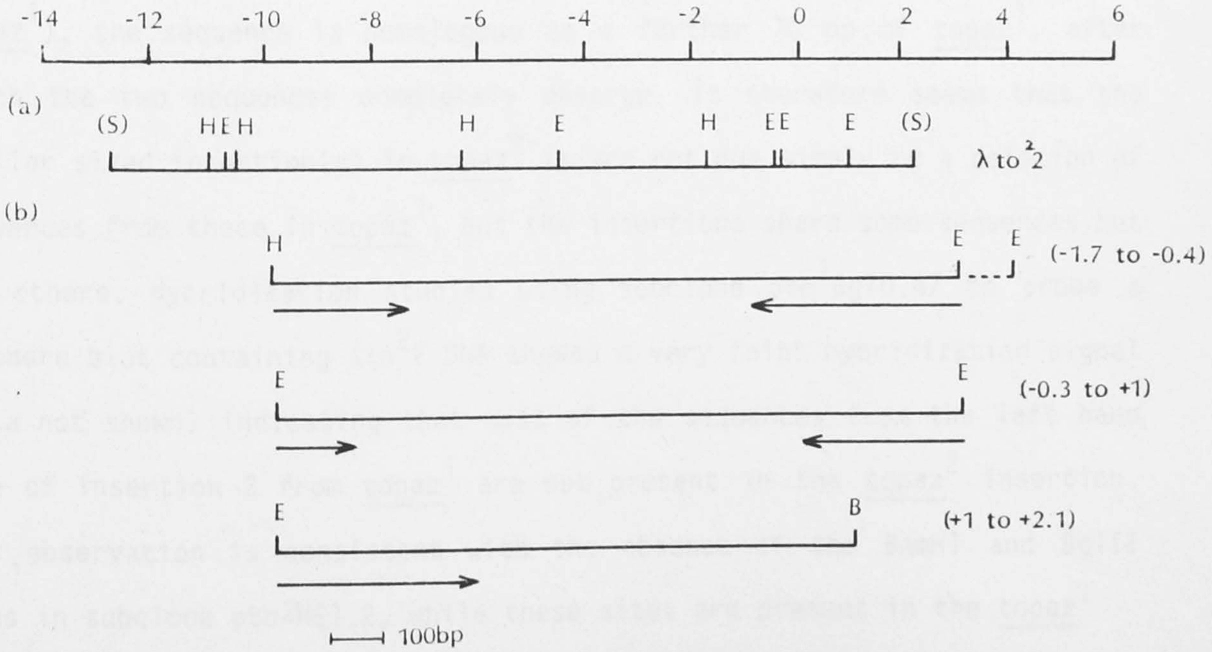
FIGURE 6.13 The sequencing strategy of the topaz^2 gene.

(a) Restriction map of clone λto^{2}2.

(b) Sequenced regions of subclones from topaz^2. Arrows indicate starting points and direction in which sequences were read from the M13 recombinants. Length of arrows indicate the extents to which sequences were read from single reactions. Restriction enzymes used: E, EcoRI; H, HindIII; (S), SalI.

The SalI sites (S) at the ends of the insert come from the poly-linker of the vector, EMBL3A.,

The presence of the EcoRI fragment (-0.4 to -0.3) has been deduced from the sequencing data (see page 100) and is marked with a broken line.



substitutions in exons; 3 are silent and one changes the amino acid glutamine to histidine (Figure 6.14). No deletions are observed in *topaz*^{2} relative to *topaz*$^{+}$; in particular the 18 bp deletion observed in exon 4 of *topaz*^{1} is not present in *topaz*^{2}. It is interesting to note that *topaz*^{2} appears to have less single base substitutions (with reference to *topaz*$^{+}$) than *topaz*^{1} does, and that out of 13 such changes in *topaz*^{2}, 6 are also present in *topaz*^{1}.

The region of the insertions in *topaz*^{2} was only partially sequenced (Figure 6.15) and again a complex pattern emerges. An additional EcoRI site was created in *topaz*^{2} by a single base substitution in insertion 2 (G to A at nucleotide position 999). This EcoRI site is only 95 bp to the left of the EcoRI site at position -275 and was not detected by the restriction enzyme analysis of clone λto^{2}2. It was found when the 1.2 kb EcoRI/HindIII fragment from *topaz*^{2} was subcloned into phage M13 and sequenced. The sequence began at a position equivalent to nucleotide 999 in insertion 2 of *topaz*^{1}. Going upstream from this EcoRI site in *topaz*^{2}, the sequence is nearly identical for about 40 bp to that of insertion 2 of *topaz*^{1}; then, allowing for small DNA insertions in *topaz*^{2} (or deletions in *topaz*^{1}), the sequence is homologous to a further 70 bp of *topaz*^{1}, after which the two sequences completely diverge. It therefore seems that the smaller sized insertion(s) in *topaz*^{2} is/are not due simply to a deletion of sequences from these in *topaz*^{1}, but the insertions share some sequences but not others. Hybridization studies using subclone pto^{1}Bgl0.47 to probe a Southern blot containing λto^{2}2 DNA showed a very faint hybridization signal (data not shown) indicating that most of the sequences from the left hand side of insertion 2 from *topaz*^{1} are not present in the *topaz*^{2} insertion. This observation is consistent with the absence of the BamH1 and BglII sites in subclone pto2HE1.2, while these sites are present in the *topaz*^{1}

FIGURE 6.14 Single base substitutions in topaz^2 exons (example).

(a) An autoradiogram of a sequencing gel showing part of the nucleotide sequence (coding strand) of exon 4 from topaz$^+$.

(b) An autoradiogram of a sequencing gel showing the nucleotide sequence (complement strand) of the same region of exon 4 as (a) only in topaz^2.

The single base substitution (nucleotide position +723 in Figure 6.2) is marked with an arrow.

This change caused the amino acid glutamine to be substituted to histidine.

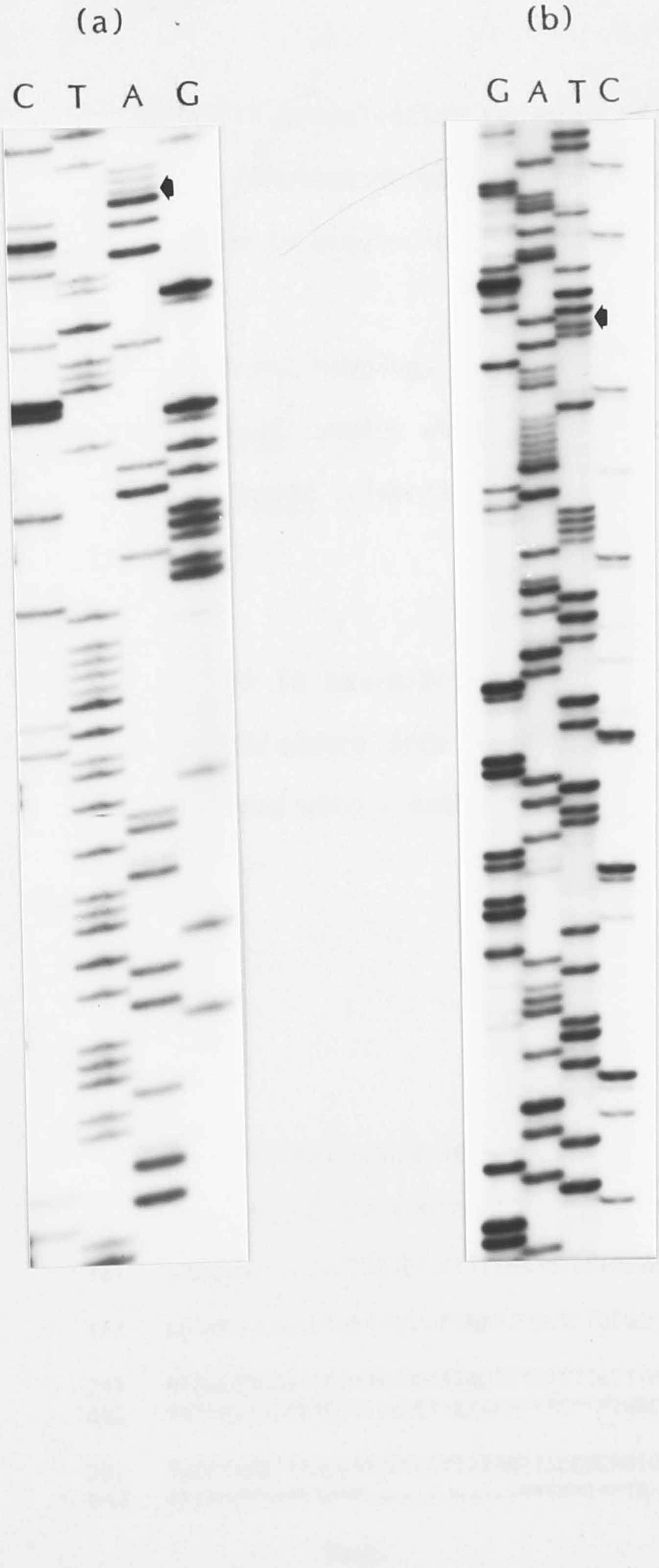

(a)
C T A G
(b)
G A T C

FIGURE 6.15 Comparison of the insertion sequences of topaz^2 with topaz^1.

The partial nucleotide sequence of the topaz^2 insertion is shown. (Nucleotide position 1 marks the first 5' nucleotide obtained by sequencing and not the first of the insertion).

Significant homology between this insertion and insertion 2 from topaz^1 begins at nucleotide position 283 in topaz^2 and 924 in topaz^1. Identical bases are marked with an asterisk *.

In order to maximize alignment of homology deletions have been introduced into the insertion 2 sequence, and these are marked with a dash -.

```
  1   ATTAAAGCAGAAATAAAAAAATGTTTCCAGCTTTCGGTTAAGATATCTCCGAAAGTGAAG

 61   TGGATCATTAAGGATCGTTATCATTTATTTTAAAGTCAACAGAATAGACTTTTATATAAA

121   GTGACAAGTGTTTTCGCTCTGTTTTTAAAACTATAGGACACGTTTTTTCCAAATCCTGGC

181   AGGAGCCCAGTTTAGAAAAATATAATTTGGTCTCTGAGTTAATTTAGTGTACCGCAAAAG

241   ATAAGCAGGTATGGAAATAAATTAATCTCTTTTAGTTACAAGACCTTTGTAAGAAGGAAA
882   *GT*ATTAA*C*ATTTTACC**C*G*A***C***TAAC*TTC****************T* insertion 2

301   TACATGAGTTTGTCATTCTCTTTGTAATTCCCACAATAGAATTTCTCAACCCTTCATATT
942   **T*********A**------------*****T**TA---T******************* insertion 2

                 EcoRI
361   ATATACATTCAGAATTC
987   *****G******G****                                            insertion 2
```

insertion. Subclone pBE9SE0.9 (that contains the 65 bp sequence homologous to the topaz^1 insertion) does hybridize to the 1.2 kb HindIII/EcoRI fragment from topaz^2, however whether this hybridization is due to the 65 bp sequence or to some other region in pBE9SE0.9 has not been established. Similarly, whether insertion 1 from topaz^1 is present in topaz^2 is not known. Both of these questions can only be answered by further sequencing of this topaz^2 region.

6.3 DISCUSSION

The comparisons of the nucleotide sequences derived from the two mutant genes, topaz^1 and topaz^2, to the topaz$^+$ gene have provided insight into three different, although related, areas. These areas are: the molecular basis for DNA sequence polymorphism in the topaz gene region, the complexity of repeated DNA sequences in L.cuprina and the difficulties in establishing the molecular nature of the topaz mutations on the basis of DNA sequencing alone.

6.3.1 Polymorphism

Restriction fragment polymorphism which is observed at the level of Southern blot analysis of genomic DNA can be caused by a number of different types of sequence changes, as illustrated in Figure 6.16.

Examples of these different types of changes have been encountered during the sequence analysis of the three topaz alleles. The observed restriction site polymorphism is due primarily to single base substitutions that have created (or abolished) restriction sites. There are two EcoRI sites and one HindIII site that are present in topaz^1 but not in topaz$^+$.

FIGURE 6.16 Molecular phenotypes of DNA Polymorphism.

Examples of changes at the DNA level (left) and the size of the bands seen on a Southern blot (right) using the region marked by |||||| as a probe and digesting the DNA with restriction enzyme H.

(a) Before any change has occurred.

(b) Sequence change abolishing an H site.

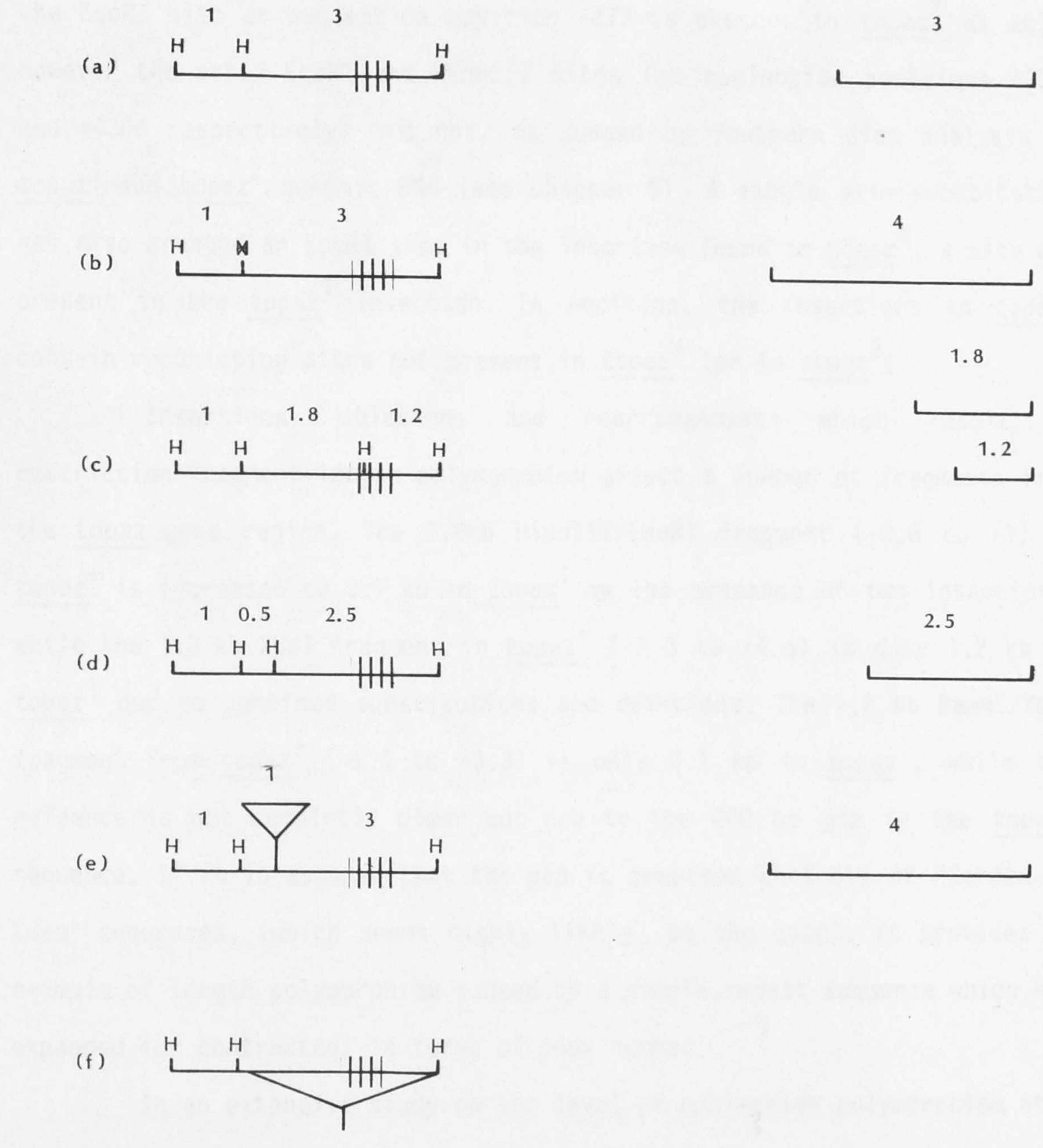
(c) Generation of a new H restriction site within the region used as a probe.

(d) Generation of a new H restriction site outside the probe region but within the H fragment detected by the probe.

(e) Insertion of a DNA sequence within the H fragment detected by the probe.

(f) Deletion of DNA sequences detected by the probe.

(g) Expansion of repeated sequences ≋ within the H fragment detected by the probe (example, hypervariable region, Jeffreys *et al.*, 1985).


1
3
H
H
H
(a)
3
1
3
H
H
H
(b)
4
1
1.8
1.2
H
H
H
H
(c)
1.8
1.2
1
0.5
2.5
H
H
H
H
(d)
2.5
1
1
3
H
H
H
(e)
4
H
H
H
(f)
1
4
H
H
H
(g)
4

The EcoRI site at nucleotide position -277 is present in *topaz*^{2} as well, however the other EcoRI and HindIII sites (at nucleotide positions +3773 and +4306 respectively) are not, as judged by Southern blot analysis of *topaz*^{1} and *topaz*^{2} genomic DNA (see Chapter 5). A single base substitution has also created an EcoRl site in the insertion found in *topaz*^{2}, a site not present in the *topaz*^{1} insertion. In addition, the insertions in *topaz*^{1} contain restriction sites not present in *topaz*$^{+}$ (or in *topaz*^{2}).

Insertions, deletions and rearrangements which result in restriction fragment length polymorphism affect a number of fragments from the *topaz* gene region. The 1.8kb HindIII/EcoRl fragment (-0.8 to +1) in *topaz*$^{+}$ is increased to 3.1 kb in *topaz*^{1} by the presence of two insertions, while the 1.3 kb XbaI fragment in *topaz*$^{+}$ (-3.3 to +4.6) is only 1.2 kb in *topaz*^{1} due to combined substitutions and deletions. The 1.2 kb BamHI/XbaI fragment from *topaz*$^{+}$ (+2.1 to +3.3) is only 0.7 kb in *topaz*^{1}; while the evidence is not completly clear cut due to the 400 bp gap in the *topaz*$^{+}$ sequence, if it is assumed that the gap is composed entirely of "Garden of Eden" sequences, (which seems highly likely be the case), it provides an example of length polymorphism caused by a simple repeat sequence which has expanded (or contracted) in terms of copy number.

In an extensive study on the level of nucleotide polymorphism at a given locus, the *alcohol dehydrogenase (Adh)* locus of *D.melanogaster* was examined (Kreitman, 1983). Eleven cloned *D.melanogaster* genes from five natural populations were sequenced. Single base substitutions as well as small deletions and insertions were observed; however length polymorphism of the type found at *topaz* (i.e., involving more than 40 bp) was not observed. The largest insertion was of 37 bp. Two of the small fragment lengths polymorphisms in *Adh* were proposed to involve expansion and contractions of homonucleotide repeats. This study showed that 1-3% change in nucleotide composition between alleles is common among populations. This is similar to that found for *topaz*, where the overall difference in nucleotide sequence between *topaz*^{1} and *topaz*$^{+}$ is about 2%.

6.3.2 Repeated sequences

The sequencing of the *topaz*$^{+}$ gene revealed the presence of a number of different repeated sequences within the gene region (see Chapter 4). In the mutant alleles, additional repeated sequences have been found, representing other types of repeats. Intron 1 has been only partially sequenced because sequences 5' to the HindIII site at nucleotide position -840 could not be subcloned from *topaz*^{1}. Repeated sequences are located in this region in both *topaz*$^{+}$ and *topaz*^{1} (as discussed in Chapter 5). Sequences from intron 1 present 3' to the above HindIII site do not contain repeated sequences in *topaz*$^{+}$, but do in both *topaz*^{1} and *topaz*^{2}. These sequences are associated with the insertions located within this region. There is no internal repetition within the insertions, at least in the case of *topaz*^{1} (data for *topaz*^{2} refers of course only to the part of the insertion that was sequenced), so their repetition in the genome has only been established by hybridization experiments (Chapter 5). The two insertions in *topaz*^{1} are located 22 bp apart, suggesting that this region in *topaz* might contain sequences that are prefered integration sites. Daniels and Deininger (1985) reported that the Alu family and other similar mammalian repetitive DNA sequence elements integrate preferentially in short AT rich regions, so the insertion of a number of elements (of the same or different types) into the same region or into each other (if they have a high AT content) is quite common. It is interesting to note that the immediate (120 bp) region surrounding the insertions is quite AT rich (about 80% in the three topaz alleles). The sequence data from the insertions, together with the preliminary analysis of other λ-clones carrying sequences homologous to them, provide a complex picture. Insertion 2 in topaz^1 is similar in part to the insertion in topaz^2. Sequences at the extreme right-hand end of the topaz^1 insertion are present (with some modifications) in topaz^2. However after only about 90 bp, the sequences diverge and remain different for the next 280 bp (up to the point to where no further sequence data is available for topaz^2). Insertion 2 of topaz^1 contains at least two different repeat units; one is the 65 bp unit, which appears to be the basis of the homology between subclone pto^{1}BE0.6 and clone λBE9 (and probably clones λBE3 and λBE6). The other repeat unit is present in the left-hand half of insertion 2 (subclone pto^{1}Bgl0.47) but its actual sequence composition has not been identified. To do this, clones carrying homologous sequences from other regions of the genome would have to be isolated and the homologous region sequenced. The two repeat units are present at about a sixty fold difference in frequency in the L.cuprina genome. These results suggest that (at the least) there has been an integration of one repeat element into another; this could have occurred at a different chromosomal location followed by transposition (as a composite) into topaz, or could have occurred in the topaz gene itself. Since the insertions in topaz^1 and topaz^2 have some sequences that are similar but others that are different, it seems likely that at least some of these complex integration events must have occurred at topaz. The obvious complexity of the insertion sequences themselves, from the point of view of the different repeat components contained within them, suggests that only the tip of an iceberg may have been revealed and a systematic examination of small fragments from within the insertion, as to their repetition number in the genome and sequence homology with other clones, would be necesssary before any conclusions about their evolution or formation can be reached.

It should be noted that a computer analysis of the insertions has not detected any direct or inverted terminal repeats that are normally associated with other transposable elements (see Chapter 1). *In situ* hybridization of these repeats to polytene chromosomes from different *L.cuprina* strains could indicate if these sequences are generally mobile in the genome, i.e., whether they are found in different chomosomal locations in different strains. If compared to other types of repeat elements found in other organisms, insertion 2 appears to be most closely similar to the "scrambled repeats" from *D.melanogaster* described by Wensink *et al*. (1979), in which clusters of repeats composed of repetitive elements arranged differently in each cluster are observed. Similar scrambled repeats have been found in the chicken (Musti *et al*., 1981), the slime mould *Physarum* (Peoples *et al*., 1985) and primates (Daniels and Deininger, 1985).

It has not yet been established whether insertion 1 is also composed of a number of repeat units that are present in different arrangements in other chromosomal locations or whether the 220 bp sequence represents the whole, or a truncated form, of one repeat element. It is clear however that insertion 1 is usually not present in close proximity to sequences from insertion 2, as probes prepared from each insertion hybridized to different recombinant phages on duplicate filters.

The other repeated sequence which is different between *topaz*$^{+}$, *topaz*^{1} (and probably *topaz*^{2} as well) is the 4 bp simple sequence found in intron 5 - the "Garden of Eden" repeat. This region appears to have diverged more rapidly than its neighbouring sequences. Apart from the apparent difference in the number of repeat units, which results in restriction fragment length polymorphism (and was already discussed in the previous section), the internal organization of TCTG and TCTA repeats is different in topaz$^+$ and topaz^1. The changes, which are by no means random but are highly constrained, mostly affect the fourth base of the repeat unit, allowing one purine to be substituted for another. The mechanism which would allow for such changes to occur is not clear. In addition, the sequence TGTA present in topaz$^+$ (as from nucleotide position 2418) in a tandem array of 35 perfect repeats until the gap in the sequence is reached, is not present in topaz^1 and might partially account for the difference in the length of the fifth intron observed between the two alleles. It seems an attractive speculation that the equivalent region in topaz^2, which was shown by Southern blot analysis of topaz^2 DNA to be much longer (3.3 kb), is also composed of the "Garden of Eden" sequence, only at a higher copy number. This suggestion, which was already proposed in the discussion of previous chapter, seems reinforced by the sequence data obtained for topaz^1.

Unequal crossing over between regions containing repeats or slippage in replication (Dover and Tautz, 1986) can perhaps explain the expansion and contraction of simple sequences, however these mechanisms do not provide a satisfying answer to the different internal organization of the repeat units. However, if the concept for the formation of scrambled clustered repeats was to be extended to these simple sequences, such that smaller blocks of the 4 bp repeat exchange positions in the genome (perhaps through homologous recombination events), two criteria would be satisfied - the length difference and the variation in the internal organization of the repeats. More topaz alleles will have to be examined before any conclusions can be reached.

6.3.3 What are the molecular explanations for the topaz mutations?

The work described in Chapters 5 and 6 has been concerned with comparisons at the level of genomic DNA, between the $topaz^+$ and the mutant alleles $topaz^1$ and $topaz^2$. Before any of the changes found will be discussed in the context of potentially causing the mutant phenotypes, the result of a preliminary Northern blot, (done with the assistance of F.V. Morris*,) should be described. The presence of the topaz transcript was detected in a number of different developmental stages in $topaz^+$ (larvae, early pupae and late pupae), and in larvae from the two mutant strains. The transcript level appeared roughly constant at the three stages examined in wild type and showed no difference in size or abundance in larvae of the three strains. This result indicates that the defects in the mutants do not affect the regulation of the gene (at least during the larval stage) and would be consistent with small changes in the DNA and consequently amino acid sequence, rather than a major rearrangement, causing the mutation.

More than one difference has been found to exist between the mutants and the wild type and none can be eliminated categorically as the source of the mutant phenotype. In intron regions, base substitutions, insertions and deletions were found. As the level and size of the transcript appears not to be altered in the mutants, changes that might result in the creation (or loss) of splice sites and affect RNA processing, or cause premature termination of transcription, can probably be ruled out. In the exons, two changes have been observed in $topaz^1$; a deletion which

* Nature of the assistance: I isolated the $polyA^+$ RNA and subcloned ptoHE1.8 into the pGEM2 vector; F.V. Morris prepared the Northern blots, the riboprobes and carried out the hybridization.

removes six amino acids and an amino acid substitution. The deletion seems to be the most likely cause of the topaz[1] mutation. Its exact effect on protein structure cannot be determined without further information about the wild type protein, however it seems likely that such a deletion could have important consequences on the structure and hence function of the topaz protein, particularly as it affects a region quite highly conserved between topaz and scarlet (Chapter 4). The other change involves a conservative amino acid substitution from isoleucine to valine. Such a change is less likely to affect the protein, particularly in light of the fact that the amino acid in the same position in scarlet is methionine; hence it appears that the amino acid at this position is not strictly conserved.

In the case of topaz[2], only one amino acid substitution has been observed, a change from glutamine to histidine. This is not a conservative substitution since histidine is a basic amino acid and glutamine is hydrophilic. Its affect on the protein function is however difficult to predict; in the case of the human factor IX protein, a change from arginine (basic) to serine (hydrophilic) was sufficient to result in a mutant phenotype due to abnormal proteolytic processing (Diuguid et al., 1986).

There are a number of difficulties involved in trying to identify the sources of the topaz mutations on the basis of the available data. The first is the lack of any information concerning exon 1 and its 5' regulatory region. Even the observation that the size and level of the topaz transcript appears unchanged in the mutants does not rule out the presence of a transposable element in that region, as the white^{sp1} mutation in D.melangoaster has an insertion 1.2 kb upstream from the transcribed region of the gene yet has no obvious affect on the level or size of the white transcript (Zachar and Bingham, 1982; Levis et al., 1984). The second difficulty is the high level of polymorphism in the *topaz* region and the absence of information about the progenitor lines from which the mutants have arisen. The fact that some of the base substitutions observed are shared between *topaz*^{1} and *topaz*^{2} indicate that they are probably not associated with the mutation itself but are part of natural sequence polymorphism. Finally, in the case of *topaz*^{2}, the failure to clone exons 6 and 7 and the limited sequence data available reduce possible interpretations.

There are a number of experiments that can help overcome these difficulties. Further work on sequencing the genes, or a more extensive Northern analysis that would examine the level of expression at all developmental stages and in different tissues, or on the cloning and sequencing of cDNA clones from the three alleles, could all be beneficial, however might still give results which were not open to completely clear-cut interpretations. The most unequivocal way to address this question is through the use of a gene transformation system and site directed mutagenesis, where specific changes can be introduced into the wild type gene, and their affect on gene function assessed *in vivo*. The use of the *topaz* and *scarlet* genes in transformation experiments will be discussed in the final chapter.

CHAPTER 7

GENERAL CONCLUSIONS AND WIDER IMPLICATIONS

The study described in this thesis was undertaken with the prospect of initiating a detailed molecular analysis of DNA sequences from L.cuprina, so as to complement the genetic and biochemical data already available for this species, as well as to provide the background for the development of a gene transformation system for the blowfly. The topaz gene was seen as a promising choice for such a study; it provided an established biochemical and genetic background and its homologous gene from D.melanogaster (scarlet) had been cloned and was suitable as a probe with which to isolate the gene.

The analysis of the topaz gene region has resulted in a number of significant advances in our understanding of gene and genome structure in L.cuprina in particular, and in dipteran insects in general, by providing important comparisons with D.melanogaster. It has revealed that the L.cuprina genome appears to be organized in a manner similar to most other genomes studied, i.e. has short repeats interspersed among the unique sequences. A high level of DNA polymorphism has been detected which is caused both by the insertion and deletion of repeated DNA sequences and by single base substitutions.

Coding regions of the topaz gene have been tentatively assigned by virtue of their homology (at the predicted amino acid level) to the deduced scarlet amino acid sequence. Interestingly, the two genes differ in their codon usage; differences have also been found in the sizes of their introns, mainly due to the presence of repeated sequences within the topaz gene. The analysis of two topaz mutants has contributed to the molecular characterization of repeated sequences and to our understanding of polymorphism in L.cuprina, even though the actual molecular explanation of the mutant phenotypes remains uncertain.

All the above observations have already been discussed in detail in the previous Chapters, and will not be dealt with further. The remainder of this Chapter will be devoted to a discussion of some of these results in the wider context of the structure/function of the topaz and scarlet proteins and of their relationship with the white protein, as well as to a description and discussion of experiments designed towards the development of a gene transformation system in L.cuprina.

7.1 Analysis of the white genes

The white gene from D.melanogaster has been cloned and studied extensively at the molecular level (see Introduction to Chapters 3 and 5) however its exact role in eye pigmentation has not yet been elucidated. As for topaz and scarlet, white is thought to be involved in the transport and storage of pigment and pigment precursors, with white being associated with both the pteridine and the ommochromme biosynthetic pathways, while topaz and scarlet are associated only with the latter.

The white gene from L.cuprina has also been cloned and sequence analysis initiated (H. Perkins, B. Wicksteed and A.J. Howells, unpublished). A partial structure of the L.cuprina white gene has been determined by sequence comparisons with the D.melanogaster white gene, using a similar approach to that described for topaz and scarlet. This analysis has shown that while there are regions of strong homology with exons from the D.melanogaster white gene (>80% at the predicted amino acid level), these are interspersed with stretches of coding region (equivalent to 9 amino acids or more) that show no homology between the genes. One possible explanation for this observation is that these regions encode parts of the protein with functions associated with the pteridine pathway. Since this pathway is different between L.cuprina (which produces the yellow pigment, sepiapterin) and D.melanogaster (which produces the red drosopterins), it might be anticipated that the white proteins will also differ.

The analysis of the structure of white is incomplete because exon 1 has not been located. In D.melanogaster exon 1 is separated by a 2.5 kb intron from the rest of the white gene (O'Hare et al., 1984) and it is possible that intron 1 will also be relatively long in L.cuprina. Sequences homologous to exons 2-6 in the D.melanogaster white have been identified in the L.cuprina white clones; a most interesting finding is that there is an additional intron in the L.cuprina gene located within the sequences homologous to exon 4 of the D.melanogaster gene. These results show that the weaker hybridization observed between the two white genes in comparison to the relatively strong hybridization signal obtained between topaz and scarlet (see Chapter 3) does not simply reflect a higher rate of divergence generally between the white genes but shows that certain regions within these genes have diverged markedly while others are as strongly conserved as topaz and scarlet are.

Other points of comparative interest to emerge from this analysis are that the introns in the L.cuprina white gene are all short (<80 bp) and contain no repeated sequences. This is unlike the situation in topaz; however it should be noted that intron 1 of white has yet to be fully defined. Simple repeated sequence DNA has been found in the white region but they lie down-stream from the final exon. Similarly, repeated sequence DNA has been found about 5 kb upstream from exon 2, in a region which might be part of intron 1 or upstream of exon 1. The codon usage of *L.cuprina white* is similar to that of *topaz*, showing a strong preference for codons ending with A or T; this is different from *D.melanogaster white* which, like *scarlet*, has a bias towards codons ending with G or C (Tearle, 1986 and see Chapter 4).

An analysis of *L.cuprina* spontaneous *white* mutants has also been carried out and the results are interesting in comparison with those presented for *topaz*^{1} and *topaz*^{2} (Chapter 6). In the case of the pure *white* null allele (*white*^{1}) an inversion, involving about 7 kb of DNA has been found. One of the inversion break points is located within the gene (near the end of exon 5) while the other is located downstream from the final exon (7); thus the inversion effectively separates the last two exons from the rest of the gene. There seems to be little doubt that the inversion provides the molecular explanation for the mutant phenotype (C. Landsberg and A.J. Howells, unpublished). Preliminary analysis of a partially pigmented spontaneous *white* mutant (*white*^{m}) has been carried out by Southern-blotting. Although the results are complicated by polymorphism (which also exists in the *white* region), no change in the restriction pattern relative to one of the major forms of the wild-type gene could be detected.

These findings, together with the results obtained for the *topaz* mutants, indicate that transposable elements might not play the same role in the generation of spontaneous mutations in *L.cuprina* as they do in *D.melanogaster*. This does not imply that such elements do not exist in *L.cuprina*, however other approaches might be needed to detect them. For example, a combined genetic and molecular biological approach would be to examine unstable mutants. There are a number of unstable mutants available in *L.cuprina* (G. Foster, pers. comm.) and these are likely candidates to contain transposable elements; however as yet there are no available probes for the isolation of these genes, so they still await molecular analysis.

7.2 Functions of topaz, scarlet and white

The DNA sequencing of the topaz and scarlet genes has provided putative amino acid sequences for the proteins and some indications about their possible structural features. Hydropathy analysis revealed that the two proteins have a short highly charged carboxy-terminus followed by a hydrophobic region; these features are typical of a membrane-spanning domain, consistent with the proteins being membrane bound and supporting the idea of topaz and scarlet being involved in transport. The putative white protein sequence also has hydrophobic regions (including a membrane-spanning structure at its C-terminus), which indicates that it too may be associated with membranes (O'Hare et al., 1984). It is, perhaps, this membrane association that explains why the white, topaz and scarlet proteins have never been identified; membrane-bound proteins are usually difficult to study since they are relatively insoluble and often present in low amounts.

In view of the similarity of their effects on the ommochrome pathway, Tearle (1986) searched for homology at the predicted amino acid level between the white and scarlet proteins. Extensive homology between the two proteins was found; if the amino acid sequences are aligned to maximize the level of homology, 49% of the amino acids are identical. These results led Tearle (1986) to propose that the white and scarlet genes are related and may have arisen by duplication of an ancestral gene, an event which must have taken place prior to the time of divergence of D.melanogaster and L.cuprina. It should be noted that the homology between scarlet and white is not evenly distributed along the genes but is concentrated in certain regions. There seems to be a good correlation between the highly conserved regions of the white genes from L.cuprina and D.melanogaster and the regions of homology between scarlet and white, indicating that these regions are probably associated with functions involving the ommochromme pathway. As a corollary, the regions which are less conserved between the white proteins and which show limited homology with scarlet are probably associated with the pteridine pathway functions. It seems quite possible that the conserved regions of the white protein on the one hand and the topaz/scarlet proteins on the other, might interact co-operatively to facilitate the transport of ommochrome pigments and pigments precursors.

Membrane-associated proteins from a number of bacterial metabolite transport systems (for example, the maltose K, phosphate B and histidine P proteins) have been studied in great detail (Higgins et al., 1982; Higgins et al., 1986; Doolittle et al., 1986; Ames, 1986) and a feature shared among some of them is the presence of two conserved sites (named A and B) along their amino acid sequences, that are thought to be involved with ATP-binding. The ATP binding sites of these membrane-bound proteins are thought to be located in cytoplasmic domains which change configuration when they bind to ATP to provide the energy for transport. S. Mount (pers. comm.) observed homology between regions in the white protein from D.melanogaster and the putative ATP binding sites of the bacterial transport proteins. These regions are also highly conserved in topaz and scarlet (Table 7.1).

Sequences homologous to site B in topaz and scarlet are located amidst the region most highly conserved between the two genes (41 amino acids of perfect homology - see Figure 4.6); it is also located in the

TABLE 7.1: THE HOMOLOGY BETWEEN WHITE, TOPAZ AND SCARLET WITH THE PUTATIVE ATP BINDING SITES OF BACTERIAL TRANSPORT PROTEINS

Consensus sequence for Site A (Ames, 1986)	G	K	S	T	L										
			*												
white (S. Mount, pers. comm.)	G	K	T	T	L										
			*												
topaz/scarlet (Chapter 4) exon 2, amino acids 3-7	G	K	T	T	L										
Consensus sequence for Site B (Ames, 1986)	E R	P	K	V	L	I L	L	D	E	P	T	S	A	L	D
	*			*									*		
white (S. Mount, pers. comm.)	D	P	P	L	L	F	C	D	E	P	T	S	G	L	D
	*			*								*	*		
topaz/scarlet (Chapter 4) exon 4, amino acids 60-74	N	P	V	I	L	F	C	D	E	P	T	T	G	L	D

* conservative substituion

region most highly conserved between scarlet and white (Tearle, 1986). Hydropathy analysis (see Figure 4.7) shows that it is located in a strongly hydrophilic region. These observations would be consistent with a hypothesis suggesting that the white, topaz and scarlet proteins are membrane-bound and that their transport functions (uptake of pigment precursors by cells and, perhaps, deposition of pigment into granules) are driven by ATP binding.

7.3 The use of the topaz and scarlet genes in P-element mediated gene transformation

The development of P-element vectors for the stable introduction of cloned genes into the germ-line chromosomes of D.melanogaster (Spradling and Rubin, 1982; Rubin and Spradling, 1982) provided a breakthrough in the study of various aspects of gene and genome structure, expression and regulation. P-element mediated transformation involves co-injecting into embryos two modified P-elements. One contains the P-element sequences recognized by the transposase (i.e., the terminal repeat regions) but not the gene coding for the transposase itself; it is into this vector that the gene to be transformed is introduced. The second (helper P) carries the functional transposase gene but lacks the inverted repeats which are essential for transposition (Karess and Rubin, 1984). Hence the former construct acts as the vehicle for carrying genes into the chromosomes while the latter provides the transposase. The development of such a system in L.cuprina would enhance not only our general molecular knowledge about the genes and genome of this species but would also contribute to the genetic program aimed at blowfly control (Cockburn et al., 1984).

Since P-element vectors have been shown to transpose in species additional to D.melanogaster (Brennan et al., 1984; Scavarda and Hartl, 1984), it seemed logical to test initially whether such vectors might function in L.cuprina. The elucidation of features of the regulation of P-element transposition in D.melanogaster (Laski et al., 1986) has revealed that transposition is restricted to the germ-line by a mechanism which operates at the level of mRNA splicing. Only in germ-line cells does the preferential splicing of one of the introns occur to give an mRNA which can be translated into functional transposase. P-element vectors have been constructed in which this intron has been removed by in vitro mutagenesis and with these somatic transposition can be obtained (Laski et al., 1986). Again, when considering the development of a cross-species transformation system, it seemed appropriate to evaluate not only conventional helper P vectors as a source of transposase but also the modified vector in case the specific splicing mechanism has been lost during evolution.

A collaborative project aimed at the development of a gene transformation system in *L.cuprina* was initiated between members of our laboratory (A.J. Howells, R.G. Tearle and myself*) and members of the CSIRO, Division of Entomology (R.B. Saint, H. Mendi, G. Clark, G.G. Foster and M.J. Whitten). A series of preliminary experiments have been carried out so far using the *topaz* and *scarlet* genes. These involved the construction of the required transformation vectors; in the case of *topaz*, the 14 kb SalI insert from clone λto8 into the SalI site of the Carnegie 4$^\#$ P-element vector and for *scarlet*, the 9kb EcoR1 fragment (-7.5 to +1.5) into the EcoR1 site of Carnegie 4. Both constructs were co-injected with either the standard P-element helper vector (pπ25.7wc) or the modified P-element helper vector pπ25.7wc Δ2-3. (This vector was generously provided by G.M. Rubin). The *scarlet* P-element construct was co-injected with Pπ25.1 into $\underline{w}^{Bwx}$; $\underline{st}^{1}$ embryos (the completely white eyes of the double mutant allows easy detection of even partial function of introduced $\underline{st}^{+}$ fragments). In one experiment 80 eggs were injected and the 13 surviving adults (16%) were mass-mated. Brown-eyed G1 progeny were recovered and were shown (by *in situ* hybridization to polytene chromosomes and by Southern blot analysis of genomic DNA from the transformed flies) to contain an additional copy of the *scarlet* gene at a new chromosomal location. The successful use of the *scarlet* gene in this transformation experiment confirms that *scarlet* can be used as a phenotypic marker.

* My own contribution: constructing the transformation vectors and performing some of the injections of *L.cuprina* embryos.

# The Carnegie 4 vector is derived from pUC and contains some P-element sequences, including the inverted repeats required for transposition (Rubin and Spradling, 1983).

Injections of topaz[1] embryos with the topaz P-element construct have been unsuccessful so far. Under injection conditions identical to those used for scarlet, the survival rate of the injected embryos was only 3%. Over 50 survivors (co-injected with the modified helper P-element) were obtained and individually mated with topaz[1] flies, however no restoration of brown pigment production could be detected among their progeny. The low survival rate and hence the small sample of flies tested prevent firm conclusions being reached from this experiment about either the use of the topaz gene or the P-element vectors.

There are a number of potential problems that need to be considered before the experiment is repeated. The first concerns the topaz gene itself. The recent finding that the first exon might be located within the SalI/XhoI fragment -7.8 to -5.1 (Chapter 4), which is located at the extreme end of the λto8 DNA used to make the construct, raises the possibility that even if the complete first exon is included within the clone, some of the required 5' regulatory sequences might not be, and a new vector might have to be constructed. Another potential problem with the topaz gene is the presence of tandemly repeated DNA sequences in introns, which might inhibit transposition. It may be possible to engineer a topaz gene which lacks these repeats by in vitro mutagenesis. One approach to testing the topaz constructs would be to use them to transform D.melanogaster st^- embryos. The presence of an additional phenotypic marker (such as neomycin resistance) would be desirable in such experiments to confirm whether transformation had occurred in cases where no restoration of eye pigmentation was apparent.

The third major potential problem in the development of a P-element mediated gene transformation system in L.cuprina concerns the use of P-element itself. It is possible that the transposase gene, which is required for the integration of the P-element construct into the genome, is not functional in *L.cuprina* because of a lack of proper recognition of regulatory sequences. Similarly, the presence of a repressor of the type present in *D.melanogaster* P-strains also cannot be ruled out (although sequences homologous to the P-element have not been found in *L.cuprina*, (A.J. Howells, unpublished). It is clear that a larger number of *L.cuprina* embryos will have to be injected and a number of different constructs tested before the usefulness of the P-element in *L.cuprina* can be fully assessed.

The discussion on transformation so far has been focussed on the practical considerations of obtaining a workable system for *L.cuprina*. It needs to be emphasised, however, that transformation experiments will have an important role to play in the investigation of gene structure and regulation in *L.cuprina*, particularly in comparison with genes from *D.melanogaster*. Especially interesting, from these respects, will be the experiments (mentioned above) involving the introduction of the *topaz* gene into *D.melanogaster*. (Hopefully the converse experiment, i.e., the introduction of the *scarlet* gene into *L.cuprina*, will also be possible in the near future).

This type of experiment can be designed to examine the conservation of regulatory sequences between the two genes and also the significance of their different codon usages; the results could be most interesting in light of the different CG composition found to exist between *topaz* and *scarlet*. Information might also be gained on the nature of the observed biochemical differences between *topaz* and *scarlet* (see Chapter 1), where *topaz* was found to be mainly active in the malpighian tubules during adult development while *scarlet* was mainly active in the eyes. A detailed analysis of the signals involved in tissue specificity between the two species might require the in vitro construction of hybrid genes consisting of parts of the regulatory region of one fused to the coding region of the other.

As outlined at the start of this Chapter, the work described in this thesis has shed light on a number of important aspects of gene and genome structure in L.cuprina. In addition, as described above, it also opens the way for further exciting work on the development of a transformation system for this species which should lead to further insights into developmental gene regulation in dipteran insects.

BLOWFLY!?
ZZZWONS!
MORE BLOWFLIES...
VRRRR...
ZZOOM!
VRRR...FFFT
VRRR...UM!
ZZWONS!
VRRR...UM!
VROOM!
ZZOOM!
©MURRONBAH
FF 12.87

-THE END-

REFERENCES

AMES, G.F. (1986). Bacterial periplasmic transport systems: structure, mechanism, and evolution. Ann. Rev. Biochem. 55 :397-425.

APPELS, R., DRISCOLL, C. and PEACOCK, W.J. (1978). Heterochromatin and highly repeated DNA sequences in Rye (Secale cereale). Chromosoma (Berl.) 70 :67-89.

ARNOLD, J.T.A. and WHITTEN, M.J. (1975). Measurement of resistance in Lucilia cuprina larvae and absence of correlation between organophosphorus-resistance levels in larvae and adults. Entomol. Exp. Appl. 18 :180-186.

BECK, T., MOIR. B. and MEPPEM, T. (1985). The cost of parasites to the Australian sheep industry. Quarterly Review of the Rural Economy 7(4) :336-343.

BEDO, D.G. (1980). C, Q and H banding in the analysis of Y chromosome rearrangements in Lucilia cuprina (Weidemann) (Diptera: Calliphoridae). Chromosoma (Berl.) 77 :299-308.

BENDER, W., SPIERER, P. and HOGNESS, D.A. (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 :17-33.

BENDER, W., AKAM, M., KARCH, F., BEACHY, P.A., PEIFER, M., SPIERER, P., LEWIS, E.B. and HOGNESS, D.S. (1983). Molecular genetics of the Bithorax complex in Drosophila melanogaster. Science 221 :23-29.

BENTON, W.D. and DAVIS, R.W. (1977). Screening λgt recombinant clones by hybridization to single plaques in situ. Science 196 :180-182.

BEVERLEY, S.M. and WILSON, A.C. (1984). Molecular Evolution in Drosophila and Higher Diptera II. A time scale for fly evolution. J. Mol. Evol. 21 :1-13.

BINGHAM, P.M. LEVIS, R. and RUBIN, G.M. (1981). Cloning of DNA sequences from the *white* locus of *D. melanogaster* by a novel and general method. Cell 25 :693-704.

BINGHAM, P.M. KIDWELL, M.G. and RUBIN, G.M. (1982). The molecular basis of P-M hybrid dysgenesis: The role of the P-element, a P-strain-specific transposon family. Cell 29 :995-1004.

BRADFIELD, J.Y., LOCKE, J. and WYATT, G.R. (1985). An ubiquitous interspersed DNA sequence family in an insect. DNA 4 :357-363.

BREATHNACH, R, and CHAMBON, P. (1981). Organization and expression of eukaryotic split genes coding for proteins. Ann. Rev. Biochem. 50 :349-383.

BREGLIANO, J.C. and KIDWELL, M.G. (1983). Hybrid Dysgenesis determinants. In: "Mobile genetic elements" (ed. J.A. Shaprio) :363-410. Academic Press. New York.

BRITTEN, R.J. and KOHNE, D.E. (1968). Repeated sequences in DNA. Science 161 :529-540.

BRENNAN, M.D., ROWAN, R.G. and DICKINSON, W.J. (1984). Introduction of a functional P element into the germ-line of *Drosophila hawaiiensis*. Cell 38 :147-151.

BUCHETON, A., PARO, R., SANG, H.M., PELISSON, A. and FINNEGAN, D.J. (1984). The molecular basis of I-R hybrid dysgenesis in *Drosophila melanogaster*: Identification, cloning, and properties of the I factor. Cell 38 :153-163.

BURR, B. and BURR, F.A. (1982). Ds controlling elements of maize at the *shrunken* locus are large and dissimilar insertions. Cell 29 :977-986.

BUSHLAND, R.C. (1971). Sterility principle for insect control. Historical development and recent innovations. In: Sterility Principle for Insect Control or Eradication. International Atomic Energy Agency, Vienna :3-14.

CHOI, O.R. and ENGEL, J.D. (1986). A 3' enhancer is required for temporal and tissue-specific transcription activation of the chicken adult β-globin gene. Nature 323 :731-734.

CLARKSON, S.G., BIRNSTEIL, M.L. and PURDOM, I.F. (1973). Clustering of transfer RNA genes of Xenopus laevis. J. Mol. Biol. 79 :411-429.

COCKBURN, A.F., HOWELLS, A.J. and WHITTEN, M.J. (1984). Recombinant DNA technology and genetic control of pest insects. Biotechnology and Genetic Engineering Reviews 2 :69-99.

COEN, E.S., THODAY, J.M. AND DOVER, G. (1982). Rate of turnover of structural variants in the rDNA gene family of Drosophila melanogaster. Nature 295 :564-568.

COLLESS, D.H. and McALPINE, D.K. (1970). Diptera. In: The Insects of Australia CSIRO Melbourne University Press :656-740.

COLLINS, M. and RUBIN, G.M. (1982). Structure of the Drosophila mutable allele, white-crimson, and its white-ivory and wild type derivatives. Cell 30 :71-79.

COLLINS, M. and RUBIN, G.M. (1983). High-frequency precise excision of the Drosophila foldback transposable element. Nature 303 :259-260.

CORNISH-BOWDEN, A. (1982). Related genes can have unrelated introns. Nature 297 :625-626.

COTE, B., BENDER, W., CURTIS, D. and CHOVNICK, A. (1986). Molecular mapping of the rosy locus in Drosophila melanogaster. Genetics 112 :769-783.

CRAIK, C.S., BUCHMAN, S.R. and BEYCHOK, S. (1981). O_2 binding properties of the product of the central exon of β-globin gene. Nature 291 :87-90.

CRAIK, C., SPRANG, S., FLETTERICK, R. and RUTTER, W.J. (1982). Intron-exon splice junctions map at protein surfaces. Nature 299 :180-182.

CRAIN, W.R., EDEN, F.C., PEARSON, W.R., DAVIDSON, E.H. and BRITTEN, R.J. (1976a). Absence of short period interspersion of repetitive and nonrepetitive sequences in the DNA of Drosophila melanogaster. Chromosoma (Berl.) 56 :309-326.

CRAIN, W.R., DAVIDSON, E.H. and BRITTEN, R.J. (1976b). Contrasting patterns of DNA sequence arrangement in Apis mellifera (Honeybee) and Musca domestica (Housefly). Chromosoma (Berl.) 59 :1-12.

DANIELS, G.R. and DEININGER, P.L. (1985). Integration site preferences of the Alu family and similar repetitive DNA sequences. Nucl. Acids. Res. 13 :8939-8954.

DAVIDSON, E.H., HOUGH, B.R., AMENSON, C.S. and BRITTEN, R.J. (1973). General interspersion of repetitive with nonrepetitive sequence elements in the DNA of Xenopus. J. Mol. Biol. 77 :1-23.

DAYHOFF, M.O., ECK, R.V. and PARK, C.M. (1972). A model of evolutionary change in proteins. In: Dayhoff M.O. (ed.) Atlas of protein sequence and structure. National Biomedical Research Foundation, Georgetown University, Washington, DC, :89-99.

DEININGER, P.L., JOLLY, D.J., RUBIN, C.M., FRIEDMANN, T. and SCHMID, C.W. (1981). Base sequence studies of 300 nucleotide renatured repeated human DNA clones. J. Mol. Biol. 151 :17-33.

DEMERS, G.W., BRECH, K. and HARDISON, R.C. (1986). Long interspersed L1 repeats in rabbit DNA are homologous to L1 repeats of rodents and primates in an open-reading-frame region. Mol. Biol. Evol. 3 :179-190.

DENNIS, E.S., DUNSMUIR, P. and PEACOCK, W.J. (1980). Segmental amplification in a satellite DNA. Restriction enzyme analysis of the major satellite of Macropus rufogriseus. Chromosoma (Berl.) 79: 179-198.

DIUGUID, D.L., RABIET, M.J., FURIE, B.C., LIEBMAN, H.A. and FURIE, B. (1986). Molecular basis of hemophilia B: A defective enzyme due to an unprocessed propeptide is caused by a point mutation in the factor IX precursor. Proc. Natl. Acad. Sci. USA 83 :5803-5807.

DOOLITTLE, W.F. and SAPIENZA (1980). Selfish genes, the phenotype paradigm and genome evolution. Nature 284 :601-603.

DOOLITTLE, R.F., JOHNSON, M.S., HUSAIN, I., VAN HOUTEN, B., THOMAS, D.C. and SANCAR, A. (1986). Domainal evolution of a prokaryotic DNA repair protein and its relationship to active-transport proteins. Nature 323 :451-453.

DOVER, G.A. and TAUTZ, D. (1986). Conservation and divergence in multigene families: Alternatives to selection and drift. Phil. Trans. R. Soc. Lond. 312 :275-289.

DOWSET, A.P. AND YOUNG, M.W. (1982). Differing levels of dispersed repetitive DNA among closely related species of Drosophila. Proc. Natl. Acad. Sci. USA 79 :4570-4574.

EDEN, F.C. (1980). A cloned chicken DNA fragment includes two repeated DNA sequences with remarkably different genomic organizations. J. Biol. Chem. 255 :4854-4863.

EFSTRATIADIS, A., POSAKONY, J., MANIATIS, T., LAWN, R., O'CONNELL, C., SPRITZ, R., DeRIEL, J., FORGET, B., WEISSMAM, S., SLIGHTOM, J., BLEECH, E., SMITHIES, O., BARALLEL, F., SHOULDERS, C. and PROUDFOOT, N. (1980). The structure and evolution of the human β-globin gene family. Cell 22 :653-668.

EMMONS, S.W., ROSENZWEIG, B. and HIRSH, D. (1980). Arrangement of repeated sequences in the DNA of the nematode Caenorhabditis elegans. J. Mol. Biol. 144 :481-500.

EMMONS, S.W. and YESNER, L. (1984). High-frequency excision of transposable element Tc1 in the nematode Caenorhabditis elegans is limited to somatic cells. Cell 36 :599-605.

EPPLEN, J.T., McCARREY, J.R., SUTOU, S. and OHNO, S. (1982). Base sequence of a cloned snake W-chromosome fragment and identification of a male-specific putative mRNA in the mouse. Proc. Natl. Acad. Sci. USA 79 :3798-3802.

ESTELLE, M.A. and HODGETTS, R.B. (1984). Insertion polymorphisms may cause stage specific variation in mRNA levels for dopa decarboxylase in Drosophila. Mol. Gen. Genet. 195 :442-451.

EVANS, B.A. (1981). Eye pigmentation in Drosophila melanogaster. Ph.D. thesis, Australian National University.

FINNEGAN, D.J., RUBIN, G.M., YOUNG, M.W. and HOGNESS, D.S. (1978). Repeated gene families in Drosophila. Cold Spring Harbor Symp. Quant. Biol. 42 :1053-1063.

FINNEGAN, D.J. (1985). Transposable elements in eukaryotes. Int. Rev. Cytol. 93 :281-325.

FLAVELL, A.J. and ISH-HOROWICZ, D. (1981). Extrachromosomal circular copies of the eukaryotic transposable element copia in cultured Drosophila cells. Nature 292 :591-595.

FOSTER, G.G. and WHITTEN, M.J. (1974). The development of genetic methods of controlling the Australian sheep blowfly Lucilia cuprina. In: The Use of Genetics in Insect Control, PAL, R. and WHITTEN, M.J. (eds.) :19-43 Elsevier, North-Holland.

FOSTER, G.G., KITCHING, R.L., VOGT, W.G. and WHITTEN, M.J. (1975). Sheep blowfly and its control in the pastoral ecosystem of Australia. Proc. Ecol. Soc. Aust. 9 :213-229.

FOSTER, G.G., WHITTEN, M.J., KONOVALOV, C. BEDO, D.G., MADDERN, R.H. and BOON, D.J. (1980). Cytogenetic studies of *Lucilia cuprina dorsalis* R-D (Diptera: Calliphoridae). Chromosoma (Berl.) 81 :151-168.

FOSTER, G.G., WHITTEN, M.J., KONAVALOV, C., ARNOLD, J.T.A. and MAFFI, G. (1981). Autosomal genetic maps of the Australian sheep blowfly, *Lucilia cuprina dorsalis* R-D (Diptera: Calliphoridae), and possible correlations with the linkage maps of *Musca domestica* L. and *Drosophila melanogaster* (Mg.). Genet. Res., Camb. 37 :55-69.

FRISCHAUF, A.M., LEHRACH, H., POUSTAKA, A. and MURRAY, N. (1983). Lambda replacement vectors carrying polylinker sequences. J. Mol. Biol. 170 :827-842.

FYRBERG, E.A., KINDLE, K.L. DAVIDSON, N. and SODJA, A. (1980). The actin genes of *Drosophila*: A dispersed multigene family. Cell 19 :365-378.

GEHRING, W.J. and PARO, R. (1980). Isolation of a hybrid plasmid with homologous sequences to a transposable element of *Drosophila melanogaster*. Cell 19 :897-904.

GERASIMOVA, T.I., MATYUNINA, L.V., ILYIN, Y.V. and GEORGIEV, G.P. (1984a). Simultaneous transposition of different mobile elements: Relation to multiple mutagenesis in *Drosophila melanogaster*. Mol. Gen. Genet. 194 :517-522.

GERASIMOVA, T.I., MIZROKHI, L.J. and GEORGIEV, G.P. (1984b). Transposition bursts in genetically unstable *Drosophila melanogaster*. Nature 309 :714-716.

GERLACH, W.L. and PEACOCK. W.J. (1980). Chromosomal location of highly repeated DNA sequences in wheat. Heredity 44 :269-276.

GILBERT, W., (1978). Why genes in pieces? Nature 271 :501.

GILLIES, S.D., MORRISON, S.L., OI, V.T. and TONEGAWA, S. (1983). A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. Cell 33 :717-728.

GOLDBERG, R.B., CRAIN, W.R., RUDERMAN, J.V., MOORE, G.P., BARNETT, T.R., HIGGINS, R.C., GELFAND, R.A., GALAU, G.A., BRITTEN, R.J. and DAVIDSON, E.H. (1975). DNA sequence organization in the genomes of five marine invertebrates. Chromosoma (Berl.) 51 :225-251.

GOLDBERG, M.L., PARO, R., and GEHRING, W.J. (1982). Molecular cloning of the white locus region of Drosophila melanogaster using a large transposable element. EMBO Journal 1 :93-98.

GRAHAM, D.E., NEUFELD, B.R., DAVIDSON, E.H. and BRITTEN, R.J. (1974). Interspersion of repetitive and nonrepetitive DNA sequences in the sea urchin genome. Cell 1 :127-138.

HAHN, S., HOAR, E.T. and GUARENTE, L. (1985). Each of three "TATA elements" specifies a subset of the transcription initiation sites at the CYC-1 promoter of Sacchromyces cerevisiae. Proc. Natl. Acad. Sci. USA 82 :8562-8566.

HARDMAN, N. (1986). Structure and function of repetive DNA in eukaryotes. Biochem. J. 234 :1-11.

HATTORI, M., KUHARA, S., TAKENAKA, O. and SAKAKI, Y. (1986). L1 family of repetitive DNA sequences in primates may be derived from a sequence encoding a reverse transcriptase-related protein. Nature 321 :625-628.

HAYSHIDA, H. and MIYTA, T. (1983). Unusual evolutionary conservation and frequent DNA segment exchange in class I genes of the major histocompatibility complex. Proc. Natl. Acad. Sci. USA 80 :2671-2675.

HEALY, M.J. (1985). Molecular and genetic studies of the uncoordinated gene of *Drosophila melanogaster*. Ph.D. thesis, Australian National University.

HEITZ, E. (1928). Das Heterochromatin der Moose. I. Jb. Wiss. Bot. *69* :762-818.

HENTSCHEL, C.C. (1982). Homocopolymer sequences in the spacer of a sea urchin histone gene repeat are sensitive to S1 nuclease. Nature *295* :714-716.

HERSHEY, N.D., CONRAD, S.E., SODJA, A., YEN, P.H., COHEN, M. Jr., DAVIDSON, N., ILGEN C. and CARBON, J. (1977). The sequence arrangement of *Drosophila melanogaster* 5S DNA cloned in recombinant plasmids. Cell *11* :585-598.

HIGGINS, C.F., HAAG, P.D., NIKAIDO, K., ARDESHIR, F., GARCIA, G. and AMES, G.F.L. (1982). Complete nucleotide sequence and identification of membrane components of the histidine transport operon of *S.typhimurium*. Nature *298* :723-727.

HIGGINS, C.F., HILES, I.D., SALMOND, G.P.C., DEBORAH, R.G., DOWNIE, J.A., EVANS, I.J., HOLLAND, I.B., GRAY, L., BUCKEL, S.D., BELL, A.W. and HERMODSON, M.A. (1986). A family of related ATP-binding subunits coupled to many distinct biological processes in bacteria. Nature *323* :448-450.

HIGGS, D.R., GOODBOURNE, S.E.Y., LAMB, J., CLEGG, J.B. and WEATHERALL, D.J. (1983). α-Thalassaemia caused by a polyadenylation signal mutation. Nature *306* :398-400.

HOLM, L. (1986). Codon usage and gene expression. Nucl. Acids. Res. *14* :3075-3087.

HUGHES, P.B. and MCKENZIE, J.A. (1986). Insecticide resistance in the Australian sheep blowfly, *Lucilia cuprina*: speculation, science and strategies. Pesticide science, in press.

HUTCHISON, C.A. III., HARDIES, S.C., PADGETT, R.W., WEAVER, S. and EDGELL, M.H. (1984). The mouse globin pseudogene βh3 is descended from a premammalian δ-globin gene. J. Biol. Chem. 259 :12881-12889.

INOUE, Y.H., and YAMAMOTO, M.T. (1986). Insertional DNA and spontaneous mutation at the white locus in Drosophila simulans. Submitted for publication.

JAGADEESWARAN, P., FORGET, B.G. and WEISSMAN, S.M. (1981). Short interspersed repetitive DNA elements in eukaryotes: Transposable DNA elements generated by reverse transcription of RNA Pol III transcripts? Cell 26 :141-142.

JEFFREYS, A.J., WILSON, V. and THEIN, S.L. (1985). Hypervariable "minisatellite" regions in human DNA. Nature 317 :67-73.

JELINEK, W.R. and SCHMID, C.W. (1982). Repetitive sequences in eukaryotic DNA and their expression. Ann. Rev. Biochem. 51 :813-844.

JOHNS, M.A., STROMMER, J.N. and FREELING, M. (1983). Exceptionally high levels of restriction site polymorphism in DNA near the maize AdhI gene. Genetics 105 :733-743.

JONES, K.W. (1970). Chromosomal and nuclear location of mouse satellite DNA in individual cells. Nature 225 :912-915.

JONES, K.W. (1983). Evolutionary conservation of sex specific DNA sequences. Differentiation 23(S) :S56-S59.

KARESS, R.E. and RUBIN, G.M. (1982). A small tandem duplication is responsible for the unstable white-ivory mutation in Drosophila. Cell 30 :63-69.

KARESS, R.E. and RUBIN, G.M. (1984). Analysis of P transposable element functions in Drosophila. Cell 38 :135-146.

KARIN, M., HASLINGER, A., HOLTGREVE, H., RICHARDS, R.L., KEROUTER, P., WESTPHAL, H.N. and BEATO, M. (1984). Characterization of DNA sequences through which cadmium and glucocorticoid hormones induce human metallothionin-IIA gene. Nature 308 :513-519.

KARLSSON, S. and NIENHUIS, A.W. (1985). Developmental regulation of human globin genes. Ann. Rev. Biochem. 54 :1071-1108.

KAUFMAN, R.J., BROWN, P.C. and SCHIMKE, R.T. (1979). Amplified dihydrofolate reductase genes in unstably methotrexate-resistant cells are associated with double minute chromosomes. Proc. Natl. Acad. Sci. USA 76 :5669-5673.

KAY, B.K. and DAWID, I.B. (1983). The 1723 element: A long, homogeneous, highly repeated DNA unit interspersed in the genome of Xenopus laevis. J. Mol. Biol. 170 :583-596.

KEDES, L.H. (1979). Histone genes and histone messengers. Ann. Rev. Biochem. 48 :837-870.

KELLER, E.B. and NOON, W.A. (1985). Intron splicing: A conserved internal signal in introns of Drosophila pre-mRNAs. Nucl. Acids. Res. 13 :4971-4981.

KREITMAN, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. Nature 304 :412-417.

KYTE, J. and DOOLITTLE, R.F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157 :105-132.

LASKI, F.A., RIO, D.C. and RUBIN, G.M. (1986). The tissue specificity of Drosophila P-element transposition is regulated at the level of mRNA splicing. Cell 44 :7-19.

LEACH, D.R.F. and STAHL, F.W. (1983). Viability of λ phages carrying a perfect palindrome in the absence of recombination nucleases. Nature 305 :448-451.

LEVIS, R., COLLINS, M. and RUBIN, G.M. (1982). FB elements are the common basis for the instability of the W^{DZL} and W^{c} *Drosophila* mutations. Cell *30* :551-565.

LEVIS, R., O'HARE, K. and RUBIN, G.M. (1984). Effects of transposable element insertions on RNA encoded by the *white* gene of *Drosophila*. Cell *38* :471-481.

LEVIS, R., HAZELRIGG, T. and RUBIN, G.M. (1986). Separable cis-acting control elements for expression of the *white* gene of *Drosophila*. EMBO Journal (in press).

LIEBERMANN, D., HOFFMANN-LIEBERMANN, B., WEINTHAL, J., CHILDS, G., MAXSON, R., MAURON, A., COHEN, S.N. and KEDES, L. (1983). An unusual transposon with long terminal inverted repeats in the sea urchin, *Strongylocentrotus purpuratus*. Nature *306* :342-347.

LINDSLEY, D.L. and GRELL, E.H. (1968). Genetic variations of *Drosophila melanogaster*. Carnegie Institute, Washington Publication :627.

LITTLE, P.F.R. (1982). Globin pseudogenes. Cell *28* :683-684.

LOHE, A.R. and BRUTLAG, D.L. (1986). Multiplicity of satellite DNA sequences in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA *83* :696-700.

LONG, E.O. and DAWID, I.B. (1979). Restriction analysis of spacers in ribosomal DNA of *Drosophila melanogaster*. Nuc. Acids. Res. *7* :205-215.

LONG, E.O. and DAWID, I.B. (1980). Repeated genes in eukaryotes. Ann. Rev. Biochem. *49* :727-764.

MADDERN, R.H., FOSTER, G.G., WHITTEN, M.J., CLARKE, G.M. KONOVALOV, C.A., ARNOLD, J.T.A. and MAFFI, G. (1986). The genetic mutations of *Lucilia cuprina dorsalis* R-D (Diptera: Calliphoridae). CSIRO Aust. Div. Entomol. Rep. No. 37 :1-40.

MANIATIS, T., FRITSCH, E.F. and SAMBROOK, J. (1982). Molecular cloning: A laboratory manual. Cold Spring Harbour Laboratory, New York.

MANNING, J.E., SCHMID, C.W. and DAVIDSON, N. (1975). Interspersion of repetitive and non-repetitive DNA sequences in the Drosophila melanogaster genome. Cell 4 :141-155.

MESSING, J., CREA, R. and SEEBURG, P. (1981). A system for shotgun DNA sequencing. Nuc. Acid. Res. 9 :309-321.

MESSING, J. (1983). New M13 vectors for cloning. Methods in Enzymology 101 :20-78.

MIKLOS, G.L.G., HEALY, M.J., PAIN, P., HOWELLS, A.J. and RUSSELL, R.J. (1984). Molecular and genetic studies on the euchromatin-heterochromatin transition of the X-chromosome of D.melanogaster: I. A cloned entry point near to the uncoordinated locus. Chromosoma (Berl.) 89 :218-227.

MILNER, R.J., BLOOM, F.E., LAI, C., LERNER, R.A. and SUTCLIFFE, G. (1984). Brain-specific genes have identifier sequences in their introns. Proc. Natl. Acad. Sci. USA 81 :713-717.

MOORE, G.P. and SULLIVAN, D.T. (1978). Biochemical and genetic characterization of kynurenine formamidase from Drosophila melanogaster. Biochem. Genet. 16 :619-634.

MOORE, G.P., SCHELLER, R.H., DAVIDSON, E.H. and BRITTEN, R.J. (1978). Evolutionary change in the repetition frequency of sea urchin DNA sequences. Cell 15 :649-660.

MONTELL, C., FISHER, E.F., CARUTHERS, M.H. and BERK, A.J. (1983). Inhibition of RNA cleavage but not polyadenylation by a point mutation in mRNA 3' consensus sequence AAUAAA. Nature 305 :600-605.

MOUNT, S.M. (1982). A catalogue of splice junction sequences. Nucl. Acids. Res. 10 :459-472.

MUSTI, A.M., SOBIESKI, D.A., CHEN, B.B. and EDEN, F.C. (1981). Repeated deoxyribonucleic acid clusters in the chicken genome contain homologous sequence elements in scrambled order. Biochemistry 20 :2989-2999.

NAGAWA, F. and FINK, G.R. (1985). The relationship between the "TATA" sequence and transcription initiation sites at the HIS4 gene of *Saccharomyces cerevisae*. Proc. Natl. Acad. Sci. USA 82 :8557-8561.

NATHANS, J., THOMAS, D. and HOGNESS, D.S. (1986). Molecular genetics of human color vision: The genes encoding blue, green, and red pigments. Science 232 :193-202.

NEWGARD, C.B., NAKANO, K., HWANG, P.K. and FLETTERICK, R.J. (1986). Sequence analysis of the cDNA encoding human liver glycogen phosphorylase reveals tissue-specific codon usage. Proc. Natl. Acad. Sci. USA 83 :8132-8136.

O'CONNELL, P. and ROSBASH, M. (1984). Sequence, structure, and codon preference of the *Drosophila* ribosomal protein 49 gene. Nucl. Acids. Res. 12 :5495-5513.

O'HARE, K. and RUBIN, G.M. (1983). Structures of P elements and their sites of insertion and excision in the *Drosophila melanogaster* genome. Cell 34 :25-35.

O'HARE, K., MURPHY, C., LEVIS, R. and RUBIN, G.M. (1984). DNA sequence of the *white* locus of *Drosophila melanogaster*. J. Mol. Biol. 180 :437-455.

ORKIN, S.H., KAZAZIAN, Jr. H.H., ANTONARAKIS, S.E., GOFF, S.C., BOEHN, C.D., SEXTON, J.P., WABER, P.G. and GIARDINA, P.J.V. (1982a). Linkage of β-thalassaemia mutations and β-globin gene polymorphisms with DNA polymorphisms in human β-globin gene cluster. Nature 296 :627-631.

ORKIN, S.H., KAZAXIAN JR. H.H., ANTONARAKIS, S.E., OSTRER, H., GOFF, S.C. and SEXTON, J.P. (1982b). Abnormal RNA Processing due to the exon mutation of β^E-globin gene. Nature 300 :768-769.

PAQUIN, C.E. and WILLIAMSON, V.M. (1984). Temperature effects on the rate of Ty transposition. Science 226 :53-55.

PARO, R., GOLDBERG, M.L. and GEHRING, W.J. (1983). Molecular analysis of large transposable elements carrying the white locus of Drosophila melanogaster. EMBO Journal 2 :853-860.

PATEL, P.I. and CASKEY, C.T. (1985). HPRT and the Lesch-Nyhan syndrom. BioEssays 2 :4-7.

PEACOCK, W.J., LOHE, A.R., GERLACH, W.L., DUNSMUIR, P., DENNIS, E.S. and APPELS, R. (1977). Fine structure and evolution of DNA in heterochromatin. Cold Spr. Harb. symp. Quant. Biol. 42 :1121-1135.

PELHAM, H.R.B. (1982). A regulatory upstream promotor element in the Drosophila hsp 70 heat shock gene. Cell 30 :517-528.

PEOPLES, O.P. and HARDMAN, N. (1983). An abundant family of methylated repetitive sequence dominates the genome of Physarum polycephalum. Nucleic Acids Res. 11 :7777-7788.

PEOPLES, O.P., WHITTAKER, P.A., PEARSTON, D. and HARDMAN, N. (1985). Structural organization of a hypermethylated nuclear DNA component in Physarum polycephalum. J. of Gen. Micro. 131 :1157-1165.

PHILLIPS, J.P. and FORREST, H.S. (1980). Ommochromes and pteridines. In: The Genetics and Biology of Drosophila. eds. Ashburner, M. and Wright, T.R.F. 2d :542-623.

PIRROTA, V., HADFIELD, C. and PRETORIUS, G.H.J. (1983). Microdissection and cloning of the white locus and the 3B1-3C2 region of the Drosophila X chromosome. EMBO Journal 2 :927-934.

POTTER, S.S., BROREIN, W.J. JR., DUNSMUIR, P. and RUBIN, G.M. (1979). Transposition of elements of the 412, copia and 297 dispersed repeated gene families in Drosophila. Cell 17 :415-427.

POTTER, S., TRUETT, M., PHILLIPS, M. and MAHER, A. (1980). Eucaryotic transposable genetic elements with inverted terminal repeats. Cell 20 :639-647.

POTTER, S.S. (1982). DNA sequence of a foldback transposable element in Drosophila. Nature 297 :201-204.

PROUDFOOT, N. (1984). The end of the message and beyond. Nature 307 :412-413.

RIGBY, P.W.J., DIECKMANN, M. RHODES, C. and BERG, P. (1977). Labelling Deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I. J. Mol. Biol. 113 :237-251.

RINCHIK, E.M., RUSSELL, L.B., COPELAND, N.G., and JENKINS, N.A. (1985). The dilute - short ear (d-se) complex of the mouse: lessons from a fancy mutation. TIG 1 :170-176.

RIO, D.C., LASKI, F.A. and RUBIN, G.M. (1986). Identification and immunochemical analysis of biologically active Drosophila P element transposase. Cell 44 :21-32.

ROEDER, G.S. and FINK, G.R. (1983). Transposable elements in Yeast. In: "Mobilegenetic elements" (ed. J.A. Shapiro) :299-328. Academic Press, New York.

ROGERS, J. (1983). Retroposons defined. Nature 301 :460.

ROLFE, M., SPANOS, A. and BANKS, G. (1986). Induction of yeast Ty element transcription by ultraviolet light. Nature 319 :339-340.

ROSE, J.K., WELCH, W.J., SEFTON, B.M., ESCH, F.S. and LING, N.C. (1980). Vesicular stomatitus virus glycoprotein is anchored in the viral membrane by a hydrophobic domain near the COOH terminus. Proc. Natl. Acad. Sci. USA 77 :3884-3888.

RUBIN, G.M. and SPRADLING, A.C. (1982). Genetic transformation of Drosophila with transposable element vectors. Science 218 :348-353.

RUBIN, G.M., KIDWELL, M.G. and BINGHAM, P.M. (1982). The molecular basis of P-M hybrid dysgenesis: The nature of induced mutations. Cell 29 :987-994.

RUBIN, G.M. (1983). Dispersed repetitive DNAs in Drosophila. In: "Mobile Genetic Elements" (ed. J.A. Shapiro) :329-361 Academic Press, New York.

RUBIN, G.M. and SPRADLING, A.C. (1983). Vectors for P element-mediated gene transfer in Drosophila. Nuc. Acid. Res. 11 :6341-6351.

SACHS, M.M., DENNIS, E.S., GERLACH, W.L. and PEACOCK, W.J. (1986). Two alleles of maize alcohol dehydrogenase have 3' structural and poly(A) addition polymorphisms. Genetics 113 :449-467.

SALSER, W., BOWEN, S., BROWNE, D., ADLI, F., FEDOROFF, N., FRY, K., HEINDELL, H., PADDOCK, G., POON, R., WALLACE, B. and WHITECOME, P. (1976). Investigation of the organization of mammalian chromosomes at the DNA sequence level. Fed. Proc. 35 :23.

SANCHEZ, F., NATZLE, J.E., CLEVELAND, D.W., KIRSCHNER, M.W. and McCARTHY, B.J. (1980). A dispersed multigene family encoding Tubulin in Drosophila melanogaster. Cell 22 :845-854.

SANGER, F., NICKLEN, S. and COULSEN, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74 :5463-5467.

SCAVARDA, N.J. and HARTL, D.L. (1984). Interspecific DNA transformation in Drosophila. Proc. Natl. Acad. Sci. USA 81 :7515-7519.

SCHMID, C.W. and JELINEK, W.R. (1982). The Alu family of dispersed repetitive sequences. Science 216 :1065-1070.

SEARLES, L.L. and VOELKER, R.A. (1986). Molecular characterization of the Drosophila vermilion locus and its suppressible alleles. Proc. Natl. Acad. Sci. USA 83 :404-408.

SEGRAVES, W.A., LOUIS, C., SCHEDL, P. and JARRY, B.P. (1983). Isolation of the rudimentary locus of Drosophila melanogaster. Mol. Gen. Genet. 189 :34-40.

SENTRY, J.W. and SMYTH, D.R. (1985). A family of repeated sequences dispersed through the genome of Lilium henryi. Chromosoma (Berl) 92 :149-155.

SHARP, P.M., TUOHY, T.M.F. and MOSURSKI, K.R. (1986). Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. Nucl. Acids. Res. 14 :5125-5143.

SINCLAIR, J.H., BURKE, J.F., ISH-HOROWICZ, D. and SANG, J.H. (1986). Functional analysis of the transcriptional control regions of the copia transposable element. EMBO J. 5 :2349-2354.

SINGER, M.F. (1982). SINEs and LINEs: Highly repeated short and long interspersed sequences in mammalian genomes. Cell 28 :433-434.

SINGER, M.F., THAYER, R.E., GRIMALDI, G., LERMAN, M.I. and FANNING, T.G. (1983). Homology between the KpnI primate and BamH1 (MIF-1) rodent families of long interspersed repeated sequences. Nucl. Acids Res. 11 :5739-5745.

SIMPSON, P.R. (1984). Molecular studies of a dispersed, simple DNA sequence in Drosophila melanogaster. Ph.D. thesis, The Australian National University.

SINGH, L., PURDOM, I.F. and JONES, K.W. (1980). Conserved sex-chromosome-associated nucleotide sequences in eukaryotes. Cold Spring Harbor Sym. Quant. Biol. 45 :805-813.

SINGH, L. and JONES, K.W. (1982). Sex reversal in the mouse (Mus musculus) is caused by a recurrent nonreciprocal crossover involving the X and an aberrant Y chromosome. Cell 28 :205-216.

SINGH, L., PHILLIPS, C. and JONES, K.W. (1984). The conserved nucleotide sequences of Bkm, which define *Sxr* in the mouse, are transcribed. Cell 36 :111-120.

SKELLY, P.J. (1985). The cuticle proteins and the cuticle protein genes of the sheep blowfly *Lucilia cuprina*. Ph.D. thesis, Australian National University.

SMITH, G.P. (1976). Evolution of repeated DNA sequences by unequal crossover. Science 191 :528-534.

SOUTHERN, E.M. (1970). Base sequence and evolution of guinea-pig α satellite DNA. Nature 227 :794-798.

SOUTHERN, E.M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. 98 :503-517.

SPRADLING, A. and RUBIN, G.M. (1981). *Drosophila* genome organization: Conserved and dynamic aspects. Ann. Rev. Genet. 15 :219-264.

SPRADLING, A.C. and RUBIN, G.M. (1982). Transposition of cloned P elements into *Drosophila* germ line chromosomes. Science 218 :341-347.

SPRADLING, A.C. and RUBIN, G.M. (1983). The effect of chromosomal position on the expression of the *Drosophila* *xanthin* *dehydrogenase* gene. Cell 34 :47-57.

STRAND, D.J. and McDONALD, J.F. (1985). *Copia* is transcriptionally responsive to environmental stress. Nuc. Acids. Res. 13 :4401-4410.

STROMMER, J.N., HAKE, S., BENNETZEN, J., TAYLOR, W.C. and FREELING, M. (1982). Regulatory mutants of the maize *AdhI* gene caused by DNA insertions. Nature 300 :542-544.

SUEOKA, M. and CHENG, T.Y. (1962). Natural occurrence of a deoxyribonucleic acid resembling the deoxyadenylate-deoxythymidylate polymer. Proc. Natl. Acad. Sci. USA 48 :1851-1856.

SULLIVAN, D.T. and SULLIVAN, M.C. (1975). Transport defects as the physiological basis for eye colour mutants of Drosophila melanogaster. Biochem. Genet. 13 :606-613.

SUMMERS, K.M. and HOWELLS, A.J. (1978). Xanthommatin biosynthesis in wild type and mutant strains of the Australian sheep blowfly, Lucilia cuprina. Biochem. Genet. 16 :1153-1163.

SUMMERS, K.M. (1979). Eye pigmentation in Lucilia cuprina. Ph.D. Thesis, Australian National University.

SUMMERS, K.M. and HOWELLS, A.J. (1980). Functions of the white and topaz loci of Lucilia cuprina in the production of the eye pigment xanthommatin. Biochem. Genet. 18 :643-653.

SUMMERS, K.M., HOWELLS, A.J. and PYLIOTIS, N.A. (1982). Biology of eye pigmentation in insects. Adv. Insect. Physiol. 16 :119-166.

SUTTON, W.D., GERLACH, W.L., SCHWARTZ, D. and PEACOCK, W.J. (1984). Molecular analysis of Ds controlling element mutations at the AdhI locus of maize. Science 223 :1265-1268.

TEARLE, R.G. (1986). Genetic and molecular biological analysis of the ommochrome biosynthetic pathway in Drosophila melanogaster. Ph.D. thesis, Australian National University.

TEARLE, R.G., BELOTE, J., McKEOWN, M.J. and HOWELLS, A.J. (1987). Cloning and characterization of the scarlet gene of Drosophila melanogaster. Manuscript in preparation.

TRUETT, M.A., JONES, R.S. and POTTER, S.S. (1981). Unusual structure of the FB family of Transposable Elements in Drosophila. Cell 24 :753-763.

ULLU, E. and TSCHUDI, C. (1984). Alu sequences are processed 7SL RNA genes. Nature 312 :171-172.

VAN ARSDELL, S.W., DENISON, R.A., BERNSTEIN, L.B. and WEINER, A.M. (1981). Direct repeats flank three small nuclear RNA pseudogenes in the human genome. Cell 26 :11-17.

VIEIRA, J. and MESSING, J. (1982). The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. Gene 19 :259-268.

WHITTEN, M.J. and FOSTER, G.G. (1975). Genetic methods of pest control. Ann. Rev. Entomol. 20 :461-476.

WHITTEN, M.J., FOSTER, G.G., VOGT, W.G., KITCHING, R.L., WOODBURN, T.L. and KONOVALOV, C. (1976). Current status of genetic control of the Australian sheep blowfly, Lucilia cuprina (Weidemann) (Diptera: Calliphoridae). Proc. XV Int. cong. Entomology, :129-139.

WELLER, P., JEFFREYS, A.J., WILSON, V. and BLANCHETOT, A. (1984). Organization of the human myoglobin gene. EMBO Journal 3 :439-446.

WENSINK, P.C., TABATA, S. and PACHL, C. (1979). The clustered and scrambled arrangement of moderately repetitive elements in Drosophila DNA. Cell 18 :1231-1246.

WALKER, A.R., HOWELLS, A.J. and TEARLE, R.G. (1986a). Cloning and characterization of the vermilion gene of Drosophila melanogaster. Mol. Gen. Genet. 202 :102-107.

WALKER, J.C., HOWARD, E.A., DENNIS, E.S. and PEACOCK, W.J. (1986b). DNA sequences required for anaerobic expression of the maize alcohol dehydrogenase-1 gene. Submitted for publication.

WEINER, A.M., DEININGER, P.L. and EFSTRATIADIS, A. (1986). Nonviral retroposons: Genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Ann. Rev. Biochem. 55 :631-661.

Wu, C. (1985). An exonuclease protection assay reveals heat-shock element and TATA box DNA-binding proteins in crude nuclear extracts. Nature 317 :84-87.

YEN, P.H. and DAVIDSON, N. (1980). The gross anatomy of a tRNA cluster at region 42A of the *Drosophila melanogaster* genome. Cell 22 :137-148.

YOST, C.S., HEDGPETH, J. and LINGAPPA, V.R. (1983). A stop transfer sequence confers predictable transmembrane orientation to a previously secreted protein in cell-free systems. Cell 34 :759-766.

YOUNG, M.W. (1979). Middle repetitive DNA: A fluid component of the *Drosophila* genome. Proc. Natl. Acad. Sci. USA 76 :6274-6278.

ZACHAR, Z. and BINGHAM, P.M. (1982). Regulation of *white* locus expression: The structure of mutant alleles at the *white* locus of *Drosophila melanogaster*. Cell 30 :529-541.