Molecular analysis of a gene affecting a quantitative trait in *Drosophila melanogaster*

Petra zur Lage

Ph.D. thesis

University of Edinburgh

1993

I declare that all the work herein was composed by myself, unless otherwise stated.

Petra zur Lage

April 1993

EX BIBL UNIV. EDINBURGH

Abstract

During an artificial selection experiment by T.F.C. Mackay for high and low abdominal bristle number in lines of *Drosophila melanogaster* which had been exposed to P element mutagenesis, a new allele of the *smooth* gene exhibiting an extremely low number of bristles on the abdomen was isolated.

The molecular cloning and characterisation of the *smooth* gene is described here. Genomic cloning of the P element insertion in the new mutant allele was carried out and sequences immediately adjacent to either side of the P element insertion were subcloned. These were used to screen a cDNA library from which a 2.6 kb cDNA clone was isolated. Northern analysis using the cDNA as a probe confirmed the presence of a single transcript of equivalent size showing varying levels of expression during development.

The 2.6 kb cDNA clone was completely sequenced and the genomic organisation of *smooth* was established, revealing that the gene consists of ten exons differing in size from 64 bp to 590 bp and introns from 60bp to more than 20 kb. The *smooth* gene extends over a genomic region more than 74 kb in length. The predicted gene product is a 52 kD protein that shows a high degree of squence similarity to a group of RNA-binding proteins and in particular to the human heterogeneous nuclear ribonucleoprotein L. The possible role of *smooth* in RNA metabolism and the way in which mutations in this gene give rise to alteration in bristle number are discussed.

Contents

Chapter 1

INTRODUCTION

1.1 Quantitative genetics

Quantitative genetics is concerned with continuous variation, and attempts to answer questions about the nature of quantitative trait genes and their inheritance. These questions are of fundamental importance not only in plant and animal breeding, but also in medicine and play a role in the understanding of the process of evolution, in particular of adaptation.

1.1.1 Quantitative characters and the underlying factors

The essential feature which defines a quantitative character is that it is continuously distributed in phenotypes. In nature, numerous examples of these characters can be observed which include any aspect of the phenotype that is measurable, ranging from anatomical dimensions and proportions (e.g. body length) to physiological functions (e.g. growth rate). When measurements of the metric characters are taken in a population, a continuous distribution of phenotypes should be observed.

The central hypothesis underlying quantitative inheritance is the segregation of alleles (polygenes) at a number of loci (quantitative trait loci) modified by environmental effects. A question which immediately arises is how many genes are involved in determining the value of a given quantitative trait. This has been a long-standing problem in quantitative genetics. The question can be addressed further by asking how many genes of large effect and how many of small effect are involved. It also has to be considered whether there is any interaction present between the genes, like linkage or epistasis, which also complicates the matter.

Estimations of the number of genes contributing to the genetic variation of quantitative characters have been obtained by two different approaches: biometrical techniques using statistical methods analysing the phenotypic variances between two populations; and secondly, by using genetic markers to map chromosome regions that contribute to the differences. In recent years, advances have been made on both the theoretical and the practical front.

Attempts are continuously being made (Zeng 1992) to improve the accuracy of the statistical predictions and estimations of the number of genes contributing to the variation of a quantitative character. The traditionally applied Castle-Wright (Castle 1921 in Hill 1984; Wright 1952) method is very basic and therefore has its limitations. In this method the difference in the means of two inbred lines is related to the variance of their F_2 and backcross populations. From the effective factor or segregation index obtained, it is impossible to distinguish what the distribution of gene effects is, i.e. how many genes involved are of major or of minor effect. Therefore one has to assume that the genes have an equal effect. The index also relies on independent gene action. But none of these are likely to apply to real characters.

The segregation of a large number of genes is not necessarily required to produce a continuous distribution of phenotypes. At the extreme, it has been shown that a normal phenotypic distribution can be obtained with the segregation at a single locus, if effects of environmental variation are taken into account (Thoday and Thompson 1976).

There is some evidence emerging that only a few major genes, i.e. 5–20, contribute to most of the phenotypic variation in a character (Shrimpton and Robertson 1988a; 1988b). Establishing the number of genes responsible for a quantitative trait requires complex segregation analysis.

Most of the early mapping experiments for trying to determine the number of genes contributing to a quantitative trait were carried out in plants. In 1909 Nilsson-Ehle studied the colour of oat glumes and the seed colour in wheat (Nilsson-Ehle in East 1910). East (1910) worked on the endosperm colour in maize and also examined the size inheritance in *Nicotiana* (East 1916). Sax (1923) studied the inheritance of seed-coat pattern and pigmentation in *Phaseolus*. In *Drosophila*, Karp (1936, in Shrimpton and Robertson 1988a) carried out the first study, attempting to determine the number of genes involved in a quantitative character, later followed by Breese and Mather (1957) and Thoday (1961). Although the work is facilitated by the markers available in *Drosophila*, it is still an immense task.

Spickett and Thoday (1966) showed that only five loci were involved in producing a 87.5% response in a line selected for high sternopleural bristle number. The most comprehensive analysis by far on the isolation of polygenic factors in *Drosophila* was carried out by Shrimpton and Robertson (1988a;1988b). In this experiment, which was an elaboration of the method used by Thoday (1961) and Spickett and Thoday (1966), the third chromosomes of two divergent sternopleural bristle lines were compared. The low line, which was the tester line, contained a recessive multiply marked chromosome and was used to analyse a high sternopleural bristle line, which had been obtained by artificial selection. The third chromosome of a high selection line was divided into five sections due to the presence of markers, which then allowed Shrimpton and Robertson (1988a; 1988b) not only to allocate the bristle effects to the chromosome sections, but also to analyse the distribution of the bristle effects within those sections. At least 17 effective factors were counted on the third chromosome contributing to the variation and some degree of epistasis between the factors was also detected. Shrimpton and Robertson (1988b) also established that the distribution of gene effects influencing the sternopleural bristle trait was clearly skewed, due to the presence of a few genes of large effects, but a larger number of genes with small effects.

With the onset of the new molecular technique of restriction fragment length polymorphism (RFLP) (Botstein *et al.* 1980), the mapping of quantitative trait loci (QTL) has been extended to organisms other than *Drosophila* e.g. tomatoes (Paterson *et al.* 1988) or maize (Edwards *et al.* 1987).

The technique relies on the observation that correlations between RFLP markers and quantitative traits can be obtained, which subsequently led (and is still being continued) to the construction of high density RFLP linkage maps for the genome of many plants and animals. With the help of these markers it has become feasible to identify individual loci for quantitative traits by cosegregation analysis. Mainly genes of large effect are being mapped, since they are in most cases of greater economic importance e.g. Booroola fertility gene in sheep (Piper and Shrimpton 1989).

Generally, it can be stated that polygenes of extremely small effect are far more difficult to identify and locate compared to genes of major effect. But the hypothesis that quantitative characters are under the control of a few major genes and a large number of minor genes wins more and more support.

1.1.2 Continuous variation and selection

What are the components of quantitative variation? Briefly summarising, the overall variation observed (phenotypic value, V_P) in a quantitative character consists of the sum of five different variance components (Falconer 1989): the genotypic value (V_G), the breeding value (V_A), the dominance deviation (V_D), the interaction deviation (V_I) and the environmental deviation (V_E). The breeding value has been labelled V_A, additive variance. It determines the degree of resemblance between relatives and V_A/V_P expresses the (narrow sense) heritability (as opposed to broad-sense heritability, which is V_G/V_P, Falconer 1989).

An important problem to address is, how is variation in quantitative traits maintained in nature? A way of examining this question is by designing selection experiments to analyse of the response of quantitative characters to selection.

The aim is to derive from the pattern of response obtained, not only information on the behaviour of a quantitative trait, but also a prediction of the distribution of gene effects.

The traditional quantitative characters to investigate in *Drosophila* include body size (measured by thorax length), wing dimensions and bristle number, in particular the sternopleural and the abdominal sternite bristles.

The two bristle characters mentioned can easily be observed under a low power microscope. When examining abdominal sternites only one of the six female sternites and one of four male sternites (the fifth male sternite does not have any bristles) are usually scored. The genetical correlation between the sternites was reported to be virtually unity (Reeve and Robertson 1954). A high heritability (i.e. with the greater part of the genetic variance being additive) had been observed in the abdominal bristle character with an environmental variance of almost zero in suitable conditions (Reeve and Robertson 1954). The sternopleural bristles, a set of bristles found on the thorax of the fly, also show a high heritability.

Examining the response of quantitative characters to directional selection, it was noticed that it is possible to change the mean phenotypic variation in some cases in only a few generations and ultimately that there is a scope to change the phenotypic variation beyond the limit, i.e. outside the range of the original variation.

Short-term (Clayton *et al.* 1957) and long-term selection (Clayton and Robertson 1957; Jones *et al.* 1969) experiments have been carried out. The response to directional selection was measured in the fourth abdominal sternite in the male and the fifth abdominal sternite in the female. Starting from a random-breeding population, Clayton *et al.* (1957) selected for five generations and noticed considerable divergence between the lines selected upwards and downwards. The response observed was attributed to the genetic variation present in the initial population, the base population.

To study the long-term effect of selection, Clayton and Robertson (1957) continued the experiment for a further 30 generations. The response obtained from the selection lines was found to be rather complex, with the different lines showing different behaviours. While in some lines a complete cessation of response to selection was found, probably resulting from homozygote lethals, other lines showed a tendency towards slowing down in the response after about 20 generations. This was thought to be due to the diminishing genetic variation present in the population. Other lines showed a continuing response to selection.

From the results obtained, it turned out to be unexpectedly difficult to make any general prediction on the response of quantitative characters to selection.

Robertson (1960) proposed the presence of a theoretical selection limit. He claimed, once a plateau in the response to artificial selection was reached, all the alleles in the population had gone to fixation and no further response was expected. From these results, he interpreted that the early variation observed was due to differences in the base population. Taking into consideration the importance of the effective population size and the selection intensity, he developed a theory of limits which allowed you to predict the limit under certain conditions. But he decided that the effect of mutation on variance had to be negligible.

Jones et al. (1968) carried out an experiment analysing the long-term response to selection under varying selection intensities and population sizes over 50 generations. Their results did not agree very well with the theory of limits to artificial selection proposed by Robertson (1960). Although in some lines the rate of response had decreased from the initial rate, showing good agreement with other selection experiments for the first generations, response to selection was still continuing at generation 50 in most of the other lines. Those latter results did not support Robertson's theory of a limit to selection caused by the loss of additive genetic variation.

In fact, it now appears that mutation must be a major force in maintaining variability (Hill 1982a; 1982b). When selection was carried out for more than 20 generations, the response was found to continue. Also contrary to Robertson's (1960) predictions, it could be shown that when selection was relaxed after several generations of directional selection, lines which were assumed to have their alleles fixed, returned to their original mean, showing a rapid response. This indicates that the cessation of response was due to the development of a balanced lethal system.

A distinction has to be made in the response to variation resulting from short-term and long-term selection. The response to short-term directional selection is predominantly due to the genetic variation present in the base population. On the other hand, the variation observed as the response to long-term selection is caused by the accumulation of spontaneous mutations.

1.1.3 Mutation and quantitative variation

Given that mutations do play a role in the response to selection, how much variation in quantitative characters is due to mutation in natural populations? Since it is experimentally more complex to study the effect of mutations on quantitative traits in natural populations, artificial selection experiments in *Drosophila* have been used to investigate this problem.

One way of examining the effect of mutations on a population is by starting off with an inbred population and waiting for spontaneous mutations to accumulate during the course of selection. Subsequently from the response obtained, it is possible to deduce how much variation is due to new mutations. This kind of experiment often requires selection to be carried out for a long period of time.

X-ray irradiation and mutation

Using a different approach, Clayton and Robertson (1955) tried to accelerate the response to variation by inducing mutations by irradiation. An inbred line was irradiated by exposing it to 1800 R of X-rays. The response to selection was followed over 17 generations. Clayton and Robertson (1955) estimated the rate of the production of new variance to occur by mutation each generation and called the number obtained mutational variance (V_m).

Although some response to irradiation was detected, the mutational variance (V_m) was not found to be significant compared to that obtained by the rate of spontaneous production of new variance each generation. Since X-rays can cause detrimental mutations, i.e. chromosome aberrations, chromosome breakage and loss, leading to sterile flies or lethality, the relatively low response detected in the irradiated lines could be attributed to that fact.

Subsequently, a number of groups (Scossiroli and Scossiroli 1959; Kitagawa 1967; Hollingdale and Barker 1971) started to investigate the effect of irradiation on quantitative characters, carrying out directional artificial selection. The results obtained varied between the groups. Increases between two-fold and one order of magnitude over the spontaneous mutation rate were reported. Although each group worked on the same quantitative character, the way the selection experiments were conducted and the units of exposure of radiation per generation induced, differed for each group. This could at least partially be the reason for the different results obtained.

Recently, it has become apparent that the spontaneous rate of mutation contributes significantly to the long-term selection response. The mutational variance (V_m) is still being used to describe the rate of genetic variation per generation produced by mutations on a quantitative character. It is scaled by the environmental variance. In *Drosophila* the approximate value for V_m was found to be $V_m=10^{-2}$ V_E. For the sternopleural and the abdominal bristle characters, the mutational variance was found to be $1.5 \times 10^{-3}V_E$ and $3.3 \times 10^{-3}V_E$, respectively (Mackay *et al.* 1992). Depending on the species and character concerned, a variety of estimates for the mutational variances have been obtained (Lynch 1988), ranging between 10^{-4} to 5×10^{-2} V_E. One explanation for obtaining this diversity of mutational variances over a range of different species is thought to be the varying generation times (Lynch 1988).

Effect of EMS on artificial selection

Another mutagenic agent is the alkylating agent ethylmethane sulfonate (EMS), the most widely used chemical mutagen in *Drosophila* mutation experiments. Although EMS is highly toxic at high concentrations, it is quite effective and shows a relatively low toxicity when an appropriate dose is administered. Mutations caused by EMS are point mutations and chromosome aberrations (contrary to earlier claims (Lim and Synder 1968 and Coté *et al.* 1986 in Ashburner 1989), there is evidence that chromosome aberrations do occur (in Ashburner 1989)).

A number of experiments on the effect of EMS on polygenic variation for viability and total fitness (Simmons *et al.* 1978 and in Mackay 1986) have been described. Simmons *et al.* (1978) established that mutants induced with EMS were partially dominant and that the mutations were found to be deleterious in male flies heterozygous for a treated second chromosome, which was consistent with the result obtained for heterozygous females by Mitchell (1977).

Reports on experiments involving quantitative characters other than fitness, i.e. the abdominal or sternopleural bristle traits, could not be traced.

When the effects of EMS are compared to those of X-ray irradiation and spontaneously occurring mutations in adult males (Gründle and Dempfle 1990), the percentage of recessive lethal mutations induced by certain doses of EMS and X-ray irradiation was equal, but 40 fold higher than in spontaneous mutation. A large difference can be observed in the number of dominant lethal mutations, where none was found in the EMS treated flies, compared to a large effect caused by X-ray irradiation. Spontaneous dominant lethals occurred three to ten times less frequently compared to the X-ray induced rate.

X-ray irradiation and EMS treatment generally only slightly increase the genetic variation available to selection. They also exhibit strong detrimental effects on fitness and, since they do represent the same distribution of mutations as compared to spontaneously occurring mutations, a different method to increase variation was required.

P element hybrid dysgenesis mutagenesis

The most recent approach to accelerating the number of mutations and therefore producing new variation in quantitative selection experiments, is by utilising the system of P element hybrid dysgenesis. A summary of the nature and mode of P transposable elements and hybrid dysgenesis will be discussed in the next section.

1.2 Transposable elements and the P element

The activity of transposable elements was first detected by Barbara (McClintock 1950), who established that there was a mechanism of gene mutation via transposable elements in maize. Only gradually it became accepted that transposable elements could not only be found in bacteria (Kleckner 1981), but that they were present in eukaryotes too, including fungi and both plants and animals. In the late 1960s the first sign of a likely presence of mutable or unstable genes in *Drosophila* was mentioned by Green (1967, 1969). In the early 1970s Hiraizumi (1971) discovered the MR element (also called the MR chromosomes), which produced mutation and mitotic recombination in a cross between a laboratory stock and a wild-type strain. Male recombination (MR) was induced when certain MR chromosomes were extracted from natural populations of *Drosophila melanogaster* and made heterozygous with a marker chromosome.

Laird and McCarthy (1968) found that 12% of the *Drosophila* genome consisted of repeated sequences with a quarter of it belonging to the groups of rRNA, 5S RNA and histone genes, while the remainder consisted of different classes or families of mobile DNA (Rubin and Spradling 1981).

It is now generally being quoted that 10-20% of the total *Drosophila* genome is made up of copies of transposable elements. More than 40 different transposable elements have been described to date (Finnegan and Fawcett 1986; Lindsley and Zimm 1992). These elements, whether they are found in *Drosophila* or any other eukaryote, can be divided into two different major classes. The first class consists of retrovirus-like mobile elements, e.g. copia-like, LINE elements and jockey, which transpose via an RNA intermediate. The second class of transposable elements are thought to transpose directly from DNA to DNA and include the P element and the foldback element of *Drosophila melanogaster* (Finnegan and Fawcett 1986; Finnegan 1989). Each class can be further subdivided according to structural differences.

The most extensively studied transposable elements in *Drosophila melanogaster* are the P family of transposable elements and the syndrome of P element hybrid dysgenesis associated with it. The latter was first described by Kidwell and Kidwell (1975; 1976) and Sved (1976).

1.2.1 Hybrid dysgenesis

Hybrid dysgenesis caused by the 'P-M' system is often associated with abnormal germ line development (Kidwell *et al.* 1977) which includes gonadal dysgenesis causing sterility, chromosome rearrangements and aberrations and mitotic recombination. The phenotypic characteristics are caused by high rates of P element transposition occurring in the germ line.

Engels (1979) realised that two components were important in regulating hybrid dysgenesis: a genetical, chromosomal component and an extrachromosomal factor, which he suggested to be maternally inherited. He introduced the term 'cytotype' indicating the presence of a cytoplasmic maternal component.

The phenomenon of hybrid dysgenesis can generally only be observed in the progeny of a cross between P cytotype males and M cytotype females, but not in the reverse cross. The cytotypes describe a regulatory state with the P cytotype being a strain, which contains P elements, but represses transposition, whereby the M cytotype or M strain does not contain any P elements, but has a permissive cellular environment for P element transposition. Therefore for mobilisation of the P element to take place, functional P elements must be introduced into the M cytotype.

The cytotype can be inherited through the female germ line in a process which closely resembles cytoplasmic maternal inheritance, i.e. following a maternal inheritance pattern. Some exceptions to the P and M cytotype rule are present.

The M' and Q strains, for example, also belong to the P family, where M' strains contain P elements, but behave as M strains during dysgenic crosses. Q strains exhibit a P cytotype, but do not show any dysgenic characteristics following crosses to either P males or M females.

P strains can be found in many natural *Drosophila* populations, carrying between 30-50 P elements scattered throughout their chromosomes (Bingham *et al.* 1982). Only one third of these P elements are complete with the remainder being different defective derivatives of the complete P element and hence showing a heterogeneous distribution in size.

Molecular information on the P element

A complete P element has a DNA sequence of 2907 bp (Rubin *et al.* 1982) with 31 bp inverted repeats on either end and 11 bp internal repeats on either side (O'Hare and Rubin 1983). In a complete P element, the four open reading frames (ORFs 0,1,2 and 3) become transcribed into a 2.7 kb primary transcript (Karess and Rubin 1984). This transcript gets processed in the germ line and after the three introns (intervening sequences IVS 1,2 and 3) are removed, the translation product obtained is a 87 kD protein coding for the transposase enzyme (Rio *et al.* 1986).

In the somatic tissue the transcript is spliced differently, retaining the third intron (IVS 3) (Rio *et al.* 1986; Laski *et al.* 1986). The resulting 2.5 kb transcript gets translated into a 66 kD truncated transposase protein, which lacks the transposase activity. This 66 kD protein is thought to be the negative regulator, repressor of transposition acting in the P cytotype.

The transposase protein has been identified to be a site-specific DNA- binding protein, which binds to a 10 bp sequence adjacent to the 31 bp inverted repeats of the P element (Kaufmann and Rio 1989). And although it has not been experimentally shown that the 66 kD repressor protein also acts as a DNA-binding protein, due to the fact that both proteins have the first 560 amino acids in common, it is assumed that the repressor protein also functions as a DNA-binding protein.

The defective P elements usually show deletions with, for example, the 1115 bp KP element (Black *et al.* 1987) having a deletion between position 808 and 2561 bp of the complete P element DNA sequence. Defective P elements are usually capable of transposition when retaining their 31 bp inverted repeats on either end, but they normally lack to ability to produce transposase.

1.2.2 Transposition

Process of transposition

The mechanism of how P elements transpose is not very well understood. It has been established that P element transposition proceeds directly from DNA to DNA without an RNA intermediate. Transposition requires a double-stranded break in the DNA and upon insertion of a P element, an 8 bp target sequence gets duplicated in the genomic DNA.

The enzyme transposase is required to catalyse P element transposition and probably also its excision. A P element is therefore not only able to insert itself (insertion of P element to a new chromosomal location is also called primary mutagenesis (Kidwell 1986)), but it is also capable of excision. In the majority of cases it does so incompletely (process also called secondary mutagenesis (Kidwell 1986)) and a change in the gene sequence can be brought about, if either some parts of the P element sequence remain inserted in the gene, or if parts of the gene get excised. In case of a complete excision the original DNA sequence is normally being restored, but mutations can also occur due to incomplete DNA repair of the double-stranded gap produced.

Although precise excision has been found to occur less frequently than imprecise excision, an increased level of precise excision has been observed (Engels *et al.* 1990) in the presence of a wild-type homologue facilitating the correct repair of the double-stranded break produced during the process of transposition.

The process of transposition as described by Engels *et al.* (1990) implies that P element transposition involves a non-replicative event. But it has also been assumed that transposition could be replicative too. These hypotheses still need to be supported by conclusive experimental data.

Imprecise excision on the other hand is due to incorrect and/or incomplete DNA repair and possibly exonuclease activity which might also be supplied by the transposase protein (Rio 1991).

Engels *et al.* (1990) and Gloor *et al.* (1991) suggested that P elements transposed via a cut-and-paste mechanism. Kaufman and Rio (1992) confirmed their suggestion by using an *in vitro* reaction system. They showed that P element transposition *in vitro* occurred by the proposed cut-and-paste mechanism and established that the transposition process required GTP as a cofactor.

Cytotype control of transposition

Although the event of transposition is very complex and only poorly understood, it has been established that transposition is regulated in two different ways: the genetically controlled system where only the progeny of a cross between a male P strain and a female M strain exhibit transposition, due to the maternal inheritance pattern of the cytotype; and secondly, a system regulated by tissue-specific restricted alternative splicing, acting on the level of RNA processing, where the P element encoded 87 kD protein transposase is only produced in the germ line, but is absent from the somatic tissue.

O'Hare and Rubin (1983) hypothesised that the P cytotype contained P factors coding for at least two functions, the transposase protein and a protein responsible for the regulation (suppression) of transposition. The regulator protein was suggested to play a role in cytotype inheritance and exert a positive feedback on its own synthesis. Before an actual repressor protein was found, different suggestions on the mechanisms of control of transposition were put forward. Simmons and Bucholz (1985) proposed that transposition was controlled by the titration of transposase, where transposase bound to defect P elements resulted in a limited amount of transposase made. They also proposed that extrachromosomal P elements were present that bound the transposase and in turn prevented any elements being mobilised by transposase.

Laski *et al.* (1986) and Rio *et al.* (1986) discovered that the P element indeed coded for a repressor protein in addition to coding for the transposase. The repressor protein was a truncated protein produced by alternative splicing. The 66kD protein was found to be present in the somatic cells, as opposed to the 87 kD transposase protein, which was only found in the germ line.

Misra and Rio 1990 showed that the repressor protein was actually present in the germ line of a P strain and suggested that the presence of a repressor during oogenesis was responsible for the maternal inheritance of P cytotype. Since the transposase protein was shown to bind to specific sequences in the P element (Kaufmann and Rio 1989), the repressor was suggested to bind the P element DNA in a similar way and hence prevent the transposase binding. The precise molecular mechanism of repression is not known yet and transcriptional, post-transcriptional or protein-protein binding mechanisms have been put forward in an attempt to explain the mechanism of repression. Kaufmann and Rio (1991) indicated that the 87 kD protein was capable of binding to transposase *in vitro*, implying its involvement in a transcriptional activation process. But *in vivo* experiments (Lemaitre and Coen 1991) did not find any support for the above observation.

On the other hand, Lemaitre and Coen (1991), while examining *in vivo* the repression of P element transposition in somatic tissues using P-lacZ fusion genes, found direct evidence for P regulatory products repressing the P promoter. They showed that the repression did not exhibit maternal effect characteristics of P cytotype in the soma. They concluded that repression by P trans-acting products was having a direct effect on the P promoter transcription, almost completely ruling out the possibility of post-transcriptional, i.e. translational and post-translational ways of regulation. P-lacZ insertion expression was repressed when present in P cytotype in all tissue or cell types, suggesting that the repressor was acting on the promoter of the P-lacZ constructs.

Siebel and Rio (1990) found an inhibitor of the 2-3 splice (the splicing of the intron between the second and third ORF) in somatic cells. This inhibitor of the 2-3 splice was also assumed to be present in the germ cells, but in smaller amounts. O'Hare *et al.* (1992) suggested that the regulation of transposition depended on the relative concentration of repressor to transposase present in the oocyte, producing a positive feedback mechanism.

The regulation/repression of transposition in strains with defect P element can be explained by the fact that these internally deleted elements only produce a truncated protein, the repressor, but not the complete transposase. A hypothesis recently proposed by Rasmusson *et al.* (1993) suggests that repression of P element transposition might also be mediated by the production of antisense P RNA which would be produced by defective P elements. This antisense RNA, which is thought to be initiated by external promoters, could subsequently bind to sense RNA molecules, hence blocking the translation of P element transcripts.

Frequency and sites of transposition

The frequency of transposition depends on several factors and varies for each transposable element. Transposable elements in general have an average rate of transposition of 10^{-4} per copy per generation (Finnegan 1989).

Regarding the P element, it has been observed that the smaller the P element, the more transposition is usually obtained. The chromosomal location of the P element and the number of P elements on the chromosomes can play a role in influencing the transposition frequency, as well as the choice of the particular P stock used. The amount of transposase produced and the activity of the repressor system are also of importance in determining the rate of transposition.

It has been shown that transposase activity can be reduced by the truncated 66 kD protein (Misra and Rio 1990), which can act as a repressor of transposase in the soma and the germ line when the source of transposase is the third chromosome integrated $\triangle$2-3 P element (Laski *et al.* 1986), which misses the third intron (IVS 3). The source of transposase in the latter construct can increase the frequency of transposition of defective elements tenfold. This elevated frequency has been explained by the fact that the modified P element has become a stable insert and no repressor protein is being produced.

A question which still needs to be answered is concerned with the specificity of P element insertion sites. Are there any hotspots? Some genes do seem to be hotspots for transposable elements, but others are not. Although P elements seem to insert throughout the *Drosophila* genome, they are usually located outside the heterochromatin (Engels 1989). P element-induced mutations show a strong tendency to be located upstream or at the 5' end of a gene or in untranslated regions of genes (Engels 1989). An 8 bp target site consensus (see chapter 4) has been compiled, but its significance or role as P element recognition site has not been established. The mechanism or any specific sites or sequence recognitions by which the P element 'decides' on where to insert has not been discovered.

For the enhancer trap method (O'Kane and Gehring 1987; Bier *et al.* 1989), where P element transformation is carried out using an especially constructed plasmid containing a weak P element promoter, the 5' end of the P element, the lacZ gene of *E.coli* and a marker gene, it has been estimated that if a saturation with P elements by mutagenesis would take place, about 50% of all *Drosophila* genes would being marked (Bellen *et al.*. 1989). It is likely that the same applies to hybrid dysgenesis, indicating that there seems to be a limit to the number of P element insertions occurring and to the number of genes being susceptible to P element insertion.

Effect of transposition on the genome

The effect of transposition can be manifold. Apart from producing deleterious mutations, transposable elements can have more subtle effects, causing different mutant phenotypes by a number of mechanisms.

The effect of a P element depends on the site of insertion with respect to the regulatory and functional regions of the gene affected. An example of where the position of transposable element insertion relative to the gene is of importance, is the *white* gene. Proximal and distal mutations have different effect on the regulation of the *white* locus expression (O'Hare *et al.* 1983).

Other cases have been reported, indicating how the gene expression of a chromosomal gene can be altered, where the insertion of a further element at a modifier locus can even suppress or enhance the effect of the initially induced mutation. The *singed* locus is a good example, where the expression of three alleles differ, dependent on the number of P elements inserted. The head to head insertion of two defective P elements produces the weakest and most unstable allele *sn^{w}* (Roiha *et al.* 1988).

The estimated frequency of mutation shows great variation dependent on the locus considered. For example, the *singed* locus has a frequency of 10^{-3} (Engels 1983) compared to that of the *alcohol dehydrogenase* gene with less than 10^{-6} (Kidwell 1986). Equally, the rate of further mutation or reversion differs. A reversion event involves an inserted P element getting mobilised in the presence of transposase in the M cytotype. Frequencies of reversions have been estimated to lie between 10^{-2} and 10^{-3} (Rubin *et al.* 1982).

Transposable elements can show a mutagenic effect for many visible and lethal loci. It has been established that the insertion of transposable repetitive elements is responsible for most of the spontaneously occurring mutations in natural populations of *Drosophila melanogaster.* Green (1988) claims that the majority of (up to 80%) of the spontaneous mutations are caused by these elements. Since transposable elements in general can be made responsible for a considerable amount of variation, they may be of evolutionary significance.

What are the consequences of transposable elements for the genome and species? Are positive or negative selection forces involved in their maintenance? The rate of replication is most likely balanced by the losses due to negative selection. It has been suggested (Charlesworth and Langley 1989), that P elements do have a negative deleterious effect on the genome, and that the number of elements present in a genome reaches an equilibrium, reducing the average fitness of the species compared to the situation prior to P element invasion. It has also been proposed that transposons could be compared to parasites increasing the risk of extinction, since laboratory experiments on *Drosophila* populations have shown that as few as eight generations' sibmating of strains containing a single P element in the first generation can lead to extinction (Engels 1992).

P element as a tool for molecular biology in *Drosophila melanogaster*

Since the discovery of the P transposable element, it has become a great tool for molecular biology, facilitating the cloning of genes by transposon tagging (Bingham *et al.* 1981). The P element can be used as a vector for gene transfer by germline transformation (Spradling and Rubin 1982). The P element enhancer trap method (O'Kane and Gehring 1987) has improved the study of tissue-specific gene expression for the understanding of developmental processes. The latest technique, developed by Gloor *et al.* (1991), involves targeted gene replacement by creating a double-stranded DNA break with the help of the P element and replacing the flanking DNA with modified sequences.

1.3 Quantitative variation and the P element

Since P element insertion has been shown to affect many visible loci, i.e. major morphological characters, Mackay (1984) considered them as ideal candidates for a tool for studying quantitative variation, anticipating that P elements also affected loci controlling quantitative variation.

The considerable advantage seen in the usage of P elements was that, not only could they be used to producing and accelerating the rate of new variation by mutation, but also subsequently, these P elements could easily be located by *in situ* hybridisation using a P element probe. The cloning of the affected gene and gene of interest would also be immensely facilitated. This could be carried out by transposon tagging (Bingham *et al.* 1981; Rubin *et al.* 1982), since a P element would be found at the site of insertion into the gene.

1.3.1 Transposable element-induced variation on quantitative traits

Mackay's first experiment (1984), studying the effects of transposition on quantitative variation by using the P transposable element, showed a dramatic result. By simply examining the effect of P element mutagenesis on the classic quantitative trait character, abdominal bristle number, she achieved in eight generations a three to six fold difference between dysgenic and non-dysgenic lines, selecting for high and low abdominal bristle number. Mackay specifically counted the number of bristles on the last abdominal sternite, which is the seventh in the female and the fifth in the male fly.

The experiment was carried out using a P strain (Harwich) and an M strain (Canton-S), setting up two replicates for dysgenic and non-dysgenic crosses with ten males and ten females each. In the subsequent generation, the bristles of 50 flies were scored for high and low abdominal bristle number. The ten highest scoring male and female flies were then used to set up the high selection line and the same was carried out for a low selection line.

The initial experiment lasting for only eight generations was continued for a further eight generations (Mackay 1985). After having carried out the first half of the experiment, Mackay (1984) already reported that some of the insertions causing mutations in the bristle number also had a deleterious effect on fitness. These lines showed an increase in the non-additive genetic variance, most possibly due to the P element insertion. This result was contrary to the belief that most of the genetic variation associated with the abdominal bristle trait was additive (Falconer 1989).

After the initial encouraging results of the first part of the experiment, during the subsequently following generations the experiment was expanded by making alterations, e.g. adding more controls and carrying out a complete analysis.

Without concentrating too much on the details of the experimental procedure (see also chapter 3 and also Mackay 1985), some of the more important additional tests and controls are mentioned below.

First of all, Mackay included two more replicate lines into the experiment. These were sublines of the Harwich (P) and Canton-S (M) strains which had been inbred for eight generations by full-sibbing. Also, starting from generation ten of the non-inbred crosses, the bristles of the last two abdominal sternites were scored in the male and female flies. Between generations ten and twelve, second and third chromosomes were extracted of each of the eight non-inbred selection lines, including the replicates.

The results obtained showed that over 16 generations on average a twofold response was achieved, comparing the dysgenic lines with the non dysgenic lines. An increase in phenotypic variance was observed in the dysgenic lines. This was shown to be due to an increase in non-additive genetic variance, with some P element induced mutations having bristle effects associated with deleterious effects on fitness. A reduced fitness was observed in the low abdominal bristle lines, since when the extracted chromosome lines were made homozygous, they became lethal.

Response to selection was found to be present on both the second and third chromosomes and there was also some evidence for it on the X-chromosome, which did not get as closely examined as the second or third chromosome. Overall, there is an indication that genes affecting bristles are present on all major chromosomes.

Overall, this experiment had shown that a great increase in new variation in the quantitative trait of abdominal bristles could be generated by P element hybrid dysgenesis. And although a large amount of P element transposition had lead to deleterious mutations, the rate of mutation creating new variation was still immensely accelerated relative to what had so far been obtained with X-ray irradiation (Mackay 1985) or EMS treatment.

This early experiment on studying quantitative variation using the tool of P element hybrid dysgenesis also had a number of weak points in the design, as mentioned by Mackay (1985) and Torkamanzehi (1989). The initial P and M strains had not been isogenic. They were unrelated and capable of producing background variation as a result. It was also suggested that the non-dysgenic crosses showed some degree of transposition and therefore had not been the perfect control to choose.

Problems for future further molecular characterisation of the location of the P element insertion, which was not carried out at that stage, were also mentioned. Apart from P element insertions (primary mutagenesis) P element excisions most probably also would have taken place, causing imprecise excision and chromosomal rearrangements (secondary mutagenesis).

Another difficulty concerning the P element insertion has to be considered, which involves the fact that usually multiple copies of P elements get inserted or are already present in a specific line. This not only decreases the chance of identifying the particular P element insertion responsible for the phenotype, but it makes it more difficult to establish that a pleiotropic phenotypic mutant is actually resulting from the insertion of one specific P element. The same pleiotropic effect on the other hand could be generated by a joint response of several P element insertions closely linked to each other on the chromosome.

Torkamanzehi *et al.* (1988) attempted to repeat the experiment using the same strains as used by Mackay (1984; 1985) and also Pignatelli and Mackay (1989) repeated the experimental procedure (and also implemented an additional experiment, i.e. using another system of hybrid dysgenesis, the I-R system, and also selecting for different bristle traits). The results obtained were quite different from what had been expected according to Mackay (1984; 1985). Torkamanzehi *et al.* (1988) could not only observe no major differences between the dysgenic and the non-dysgenic selection results, but also obtained a low non-dysgenic line actually showing a large response. Transposition was suggested to have occurred in the non-dysgenic lines, rendering those unsuitable for controls.

A different possible explanation for obtaining such large response in the initial selection experiments by Mackay (1984; 1985) was suggested by Engels (1989), who pointed out that the original, preexisting P element sites in the Harwich P strain could have contributed to the effect.

A major investigation (Shrimpton *et al.* 1990) was carried out determining and analysing the P element insertion sites in each of eight dysgenic and non-dysgenic lines in a continuation of the Mackay experiment (1985), using *in situ* hybridisation. It became established that new insertions of P elements had taken place. This was not only the case for the dysgenic lines, but also for the non-dysgenic lines. Clearly, here was the evidence for the suspicion that P element transposition did occur in the non-dysgenic lines. The transpositions in the non-dysgenic lines were explained by the fact that the repression system in the non-dysgenic F1 hybrids had failed, because they had been mated with each other for several generations beforehand (Lai and Mackay 1990).

1.3.2 Other work on selection involving P elements

Since the early attempts by Mackay, a small number of other groups (e.g. Torkamanzehi *et al.* 1988; 1992) apart from Mackay (1987; Pignatelli and Mackay 1989; Lai and Mackay 1990), have been studying the effect of P elements on other quantitative characters. The effect of P element transposition on fitness, with the two major components of viability and fecundity, was also examined (Mackay 1985; 1986; 1989; Fitzpatrick and Sved 1986; Eanes *et al.* 1988).

Different approaches were applied, using for example chromosome contamination experiments (Mackay 1986; 1987; Lai and Mackay 1990) by passing an M strain through non-dysgenic and dysgenic crosses for a single generation and determining its effect on mutational variance by comparing the non-dysgenic with the dysgenic lines (and showing that they were almost equal in their response).

Improvements have been made with regard to the difference between the initial P and M strains, trying to reduce it as much as possible and therefore aiming at obtaining isogenic lines and decreasing the level of background mutations in the base population (Torkamanzehi *et al.* 1992). Other advances have become possible due to the development of the especially designed third chromosomes containing the modified P[ry$^+$; $\triangle$2-3](99b) P element (Robertson *et al.* 1988; Mackay 1992a), a constant source of transposase.

On the whole, even though the controls in the initial experiment did not turn out to be suitable, the outcome of the experiment nevertheless proved that P element mutagenesis can be a powerful method to increase mutational variance by varying degrees (i.e. depending on the experimental design). Variation had been shown to be due to P element transposition and not resulting from background mutations caused by spontaneous mutation. The experiments indicated that accumulation of P element insertions at quantitative trait loci can change the phenotype of that trait, when under conditions of directional selection (usually carried out for both directions).

The mutation rate obtained by hybrid dysgenesis had been shown to exceed that of spontaneously occurring mutations, even though P elements do have a detrimental effect. Mackay (1986) suggested that the mutations obtained for quantitative variation might even mimic those of natural populations and therefore indicated that P element hybrid dysgenesis might be an ideal system to work on, exhibiting a close resemblance to the natural system of spontaneous mutations. Whether this is the case still has to be shown. Many natural occurring mutations are point mutations and small insertions (Langley *et al.* 1988), whereas point mutations can normally not be produced by P elements. Nevertheless, the method of using P elements to study quantitative variation is not impaired by that fact.

Due to the design of the experiments, predominantly major phenotypic effects on quantitative traits are being observed using P element hybrid dysgenesis. Therefore can these experiments contribute to the classical question on trying to understand how many genes are involved in a quantitative character and what is the distribution of the gene effects? Although there is a clear bias towards finding genes of major effect, it might also be emphasising and supporting the mostly accepted hypothesis that there are a few genes of large effect and a large number of genes of small effect contributing to the quantitative character.

Mackay (1992a; 1992b) produced evidence showing that the distribution of P element effects on the metric characters for abdominal and sternopleural bristles in artificial selection experiments was asymmetric (skewed) and leptokurtic. These results agree with those collected by experiments using direct data (Yoo 1980; Caballero *et al.* 1991; Santiago *et al.* 1992), i.e. accumulation of spontaneous mutations without the application of a mutagen, and those of Shrimpton and Robertson (1988a; 1988b) in their mapping experiment. The pleiotropic effects of mutations on fitness (Mackay 1992; Santiago *et al.* 1992) was also shown to be skewed and highly leptokurtic.

1.3.3 Nature of quantitative trait genes

One question which has not yet been addressed in the above is, whether any quantitative trait loci have actually been identified either by cytological or physical mapping or even molecularly by analysing the point of P element insertions in *Drosophila.* And if so, has it been determined what these genes are?

In the RFLP analysis determining QTLs, several traits have at least been physically mapped, for example in maize (Doebley and Stec 1991) and tomato (Paterson *et al.* 1988; 1990). Paterson *et al.* (1988) for example succeeded in mapping several QTLs for quantitative characters controlling fruit mass, soluble solids and pH in tomato. All of these were identified as QTLs of relatively large effect and therefore possible useful information for breeding purposes.

What kind of genes are quantitative trait genes? Presently there have mainly been a number of speculations on the nature of quantitative trait genes without any definite conclusive evidence.

Thompson (1975) suggested that polygenes might directly act as a suppressor or enhancer controlling the mutant locus. Mukai and Cockerham (1977) supported that theory by indicating that polygenes might be located outside the structural genes on the chromosome in what they called a 'controlling' region. On the other hand, Thompson (1975) also mentioned that polygenes could primarily be involved in some developmental process and hence only indirectly modify the expression of a mutation in that process.

Mackay (1985; 1992a) also made several suggestions on the nature of quantitative trait genes. She proposed that maybe the alleles responsible for quantitative genetic variation could have major effects on related traits. Another possibility she suggested was that loci, affecting quantitative characters, could have a range of allelic effects. Those of large effect could then be described as the major genes.

In a small number of experiments selecting for the abdominal bristle number in *Drosophila*, a few major mutations of major effect have been identified. Three of these genes were *scabrous* (Jones *et al.* 1968), *bobbed* (Frankham *et al.* 1978; Frankham *et al.* 1980) and *scute* (Yoo 1980).

Scabrous, which has been found to have a large effect on abdominal bristle number in both homozygotes and heterozygotes, increases the bristle number. *Scabrous* has independently appeared several times in different selection lines (Jones *et al.* 1968).

Molecular analysis of the *scabrous* gene, which also affects the eye development by irregular spacing of ommatidia, has recently been identified to function as a regulator in *Drosophila* neurogenesis (Mlodzik *et al.* 1990).

Mutant alleles at the *bobbed* locus also arose several times independently in selection experiments for bristle number. The 18S and 28S ribosomal RNA subunits (Ritossa 1976) are encoded at this locus. Further investigations showed that the *bobbed* mutation had been caused by an unequal cross-over event, producing a partial deficiency at the locus (Frankham 1978). Frankham (1989) has also confirmed the appearance of a spontaneous *bobbed* mutation in Clayton and Robertson's selection experiment (1957).

Yoo (1980) indicated that he had come across a genuine spontaneous mutation of *scute* (see below for function of *scute*) during his long-term selection experiment for abdominal bristle number, since attempts failed to identify a *scute* allele segregating in his base population. He also noticed a *scabrous* mutation in his experiment, but could not satisfactorily trace the origin of that mutation to the base population.

1.4 The *smooth* discovery

The initial selection experiments carried out by Mackay (1984; 1985) using P element hybrid dysgenesis, led to the discovery of a gene of major bristle effect, the *smooth* gene.

1.4.1 The origin of *smooth*

As mentioned above, second and third chromosomes were extracted between generations ten and twelve from the dysgenic high and low lines. It was determined that half of the extracted third chromosomes of the high selection line were homozygous lethal and possibly being selected for due to their heterozygous effect of increasing bristle score. No further attempt to map these third chromosomes was made.

On the other hand, a second chromosome extracted from the low dysgenic selection line exhibited a major bristle effect showing an extreme phenotype. Almost all of the abdominal sternite bristles were missing and further examination of mutants disclosed that also other microchaetae were lost or reduced, e.g. the acrostichal hairs, microchaetae around the eye and the aristae. It was also observed that some of the scutellar macrochaetae were thinner than normal. The reduced viability of this low selection line was thought to be associated with this abnormal bristle allele.

Further examination showed the allele to be semi-lethal (although it should more appropriately be described as lethal, showing a frequency of less than 3% in a population), recessive and therefore not maintainable as a homozygote, and was female sterile (Mackay 1985). As described in more detail in chapter 3, the allele was subsequently mapped by Professor A. Robertson. Initially, it was thought to be located at the approximate position of 2-80 by physical mapping on the right arm of the second chromosome.

Since the phenotypic description of a *smooth* mutation, found in Lindsley and Grell (1968), and also the approximate location complied with that of the newly obtained mutation, a complementation experiment with a *smooth* allele (acquired from the Bowling Green stock centre) confirmed the hypothesis that a new allele of the *smooth* locus had been isolated.

After having mapped the mutation, it still had to be established that a P element insertion had caused the mutation. Therefore, the cytological positioning of a P element associated with the new *smooth* allele, *sm*^{3}, was subsequently carried out by Dr. A. Shrimpton (see also chapter 3). With a combination of second chromosome extraction, complementation and *in situ* hybridisation, he identified, out of a cluster of six P elements in the region of interest, one particular P element at 56 E to be responsible for the mutation.

To further ensure that the P element identified was the correct one, Shrimpton carried out a physical mapping of the P element by scoring the recombinants between the line containing the P element at 2-91.5 and a marked second chromosome (see chapter 3). The results obtained confirmed the position of the P element.

Although it had now been established that the P element had caused the mutation, it was not necessarily evident from the experiment that the *smooth* mutation had arisen *de novo* during the course of selection. One possibility to be considered would be that the allele could have been segregating in the Harwich population at the start of the experiment. The P element insertion at 56E had been found in three independent low lines showing phenotypes similar to *smooth*, suggesting that the allele could have been segregating in the Harwich population at the start of the experiment. Since no *in situ* hybridisations detecting P elements had been carried out on the original Harwich strain, no direct evidence for this suggestion could be supplied.

Another point which has to be mentioned is that P element insertions detected as present in one polytene chromosome band do not necessarily imply that the insert is at the same position as a gene, since a chromosome band contains on average 30 kb of DNA. Therefore, although P elements might be present in the other lines on the 56E polytene chromosome band, the actual insertion might not be at precisely the same position. *In situ* hybridisation would not be sensitive enough to show a distinction in that case. Further, the particular site of the *smooth* locus could be a hotspot for P element insertion and the P elements could be inserted at the same position. Since no comparisons at the molecular level of the P element location for the other three lines were carried out, the answer to this question is not known.

The analysis of the *smooth* locus was continued and Shrimpton carried out crosses in order to obtain reversions of the phenotype. He reported two cases of precise P element excision, as shown by *in situ* hybridisation (and later established by Southern blots), associated with the change to wild-type.

Subsequently, Shrimpton succeeded in cloning the area of P element insertion on the second chromosome by P element homology mapping (see chapter 4) covering a region of approximately 50 kb. He found that a full-length P element had been inserted, which was later confirmed by DNA sequence analysis (see chapter 4).

The other *smooth* alleles Two *smooth* alleles existed prior to the one described above. The first *smooth* (*sm*) allele had been discovered by Bridges and Brehme (1944) and had been mapped at 2-91.5. The phenotypes of *smooth*(*sm*) and the new *smooth* (sm^3) closely resemble each other (see chapter 3).

A second *smooth* allele had been found by Frankham and Nurthen *et al.* (1981). This allele, originally called $smooth^{lab}$, (sm^{lab}), but renamed sm^2 (Lindsley and Zimm 1990), appeared in a Canberra outbred population during a selection experiment. The phenotype was found not to be as severe as in the other two mutants (see chapter 3).

1.5 Development of bristles in *Drosophila*

1.5.1 Description of development of abdominal bristles

The reason for studying quantitative variation on the abdominal bristles was mentioned above. But how do abdominal bristles and bristles in general develop?

When examining an adult fly under the microscope, one immediately notices that the insect is covered in hairs, ranging from macrochaetes (large bristles) to microchaetes (small bristles), which can be mechanosensory or chemosensory (e.g. bristles on the wing margins) and trichomes (cell hairs).

The bristle pattern is very characteristic in *Drosophila* and can also be found among the higher Diptera. The macrochaetes show a constant number and precise positioning, while the evenly distributed microchaetes vary only slightly in their number and positioning. Due to its constancy and the fact that the macrochaetes are individually identifiable (also called landmark bristles), the bristle pattern has been used for taxonomic purposes. It has been discovered that the bristle patterns have been present for well over 50 million years, which would represent more than 100 million generations according to Sturtevant (1970).

In the following, the development of the abdomen and its bristles from the embryo to the adult is briefly summarised.

Just after blastoderm formation in *Drosophila* embryogenesis, i.e. as early as approximately 3 h after egg laying, small groups of cells are selected and set aside to form adult structures later in development. These clusters of 10–50 neighbouring cells are distinct, although not visibly, from the cells that differentiate during embryogenesis to form larval tissues. They will form the imaginal discs and histoblast nests during larval development, which contain the anlagen for adult structures.

The imaginal discs and histoblast nests do not have any larval role, remaining diploid and undifferentiated during larval development, while other cells become polyploid and are involved in the formation of larval structures.

There are nine major pairs of imaginal discs in the cephalic and thoracic region of the larvae, plus an additional single genital disc in the genital region. These discs plus the histoblast nests construct the whole adult integument.

The histoblast nests are segmentally repeated in the larval abdomen and consist of an average of ten imaginal cells per nest. They are positioned amongst the larval epidermal cells, which they eventually replace (during the pupal stage) and subsequentially form the adult abdominal epidermis (Madhavan and Madhavan 1980).

Whereas the imaginal discs proliferate and multiply during the larval period independently from each other, resulting in different shape and forms, for example the wing and the eye disc grow exponentially during the larval development (Garcia-Bellido and Merriam 1971), the histoblast imaginal cells remain constant during embryonic and larval development. They only start growing rapidly three hours after pupariation (120 h after egg laying at 25°C) and continue growth up to 15 h after pupariation. Up to this stage, they do not replace any larval cells and remain confined to their original area.

From 15 h to 36 h after pupariation, the histoblast nests enlarge and gradually replace the polytene larval epidermal cells with the adult epidermis (Roseland and Schneiderman 1979). Apart from the secretion of the cuticle, differentiation of the histoblasts also leads to the formation of tendons and muscle attachment sites.

Each adult abdominal segment originates from six 'histoblast nests (a pair of each posterior dorsal, anterior dorsal and ventral nests) and two spiracular anlagen, which are present in the equivalent position of the larvae.

The anterior and posterior dorsal nests give rise to the presumptive anterior and posterior hemitergite. The hemitergites eventually fuse along the midline to form the adult tergite. The ventral histoblast nests are responsible for the formation of the adult ventral surface, i.e. the sternites, which are small plates located along the ventral midline. The tergites are separated from the sternites by the pleural membrane within which, close to the lateral margins of the tergites, are the spiracles. There is one spiracle per hemitergite (Roseland and Schneiderman 1972).

In the adult female, fly the dorsal surface of the abdomen consists of eight segments, the tergites, seven of which are derived from the histoblast nests and the eighth segment is formed by the genital imaginal disc. The sternites are labelled from two to seven in the female, with the first sternite missing.

The male abdomen is smaller than the female abdomen and has neither a seventh tergite nor a seventh sternite. The sixth sternite is bristleless and the sixth tergite is wider than that of the female and has two lateral spiracles instead of one.

The male and female abdominal pigmentation differs. In the female, the posterior one-third of each tergite shows black pigmentation. The fifth and the sixth tergites are completely pigmented in the male.

The tergites exhibit a constant bristle pattern. Most of the tergites are covered by trichomes (cell hairs), with the microchaetes and macrochaetes arranged in rows (three rows of microchaetes followed by one row of macrochaetes in the anterior–posterior direction, Roseland and Schneiderman 1979).

The sternites can be described as little rectangular plates along the midline of the ventral abdominal surface. They are covered with evenly spaced macrochaetes and microchaetes. No trichomes are present and there is no pigmentation (Santamaria and Garcia-Bellido 1972; Bryant 1978). The abdominal pleura does not contain any bristles.

A number of researchers (Sondhi 1964; Claxton 1974) have attempted to establish an order in the positioning of the sternite bristles. Sondhi (1964) tried to establish a pattern in the arrangement of the bristles on the fourth sternite, in particular showing an interest in those bristles with an anteroposterior direction. He suggested the presence of an axial gradient in the anterior to posterior orientation, due to the fact that he managed to show a linearly diminishing response from the anterior to posterior sternites in the selection experiment.

Claxton (1974) was interested, as Maynard Smith and Sondhi (1961) before, in answering the question of the existence or non-existence of rows of bristles on the sternite. Claxton came to the conclusion by just examining the bristles of the third sternite that there was no evidence for a specific mechanism placing bristles in rows. He suggested that there is a constancy of some sternite bristle sites. He found a high frequency in the positioning of large bristles in each posterior lateral corner of each sternite.

As far as I am aware, no further work has been carried out on the sternites asking Sondhi's and Claxton's questions. But, generally speaking, the sternite chaetes do show variation in their number, positioning and orientation, but have a tendency towards an even spacing.

1.5.2 The development of the nervous system

The chaetes (macro- and microchaetes) belong to the external sensory class of sensilla in *Drosophila.* Other sensilla are the internal sensory organs, e.g. the chordotonal organs which function as strain receptors and the multiple dendrite organs. The external and internal sensory organs are part of the peripheral nervous system.

Very early during embryonic development (immediately after gastrulation and during germ band elongation), a part of the ventral ectoderm will give rise to the neurogenic region, the neuroectoderm. Cells of the neuroectoderm segregate at approximately 5 h after egg laying and will form either the neural lineage, producing neuroblasts as part of the presumptive central nervous system (CNS), or the epidermal lineage, producing epidermoblasts, the progenitor of the epidermis or sensory organ progenitor cells, as part of the peripheral nervous system (PNS) (Ghysen and Dambly-Chaudiere 1989; Jan and Jan 1990). Proneural genes and neurogenic genes are acting on the neuroectodermal cells in order to produce the two different lineages, neural or epidermal, developing into the CNS and PNS, respectively.

The proneural genes are the four genes (*achete*, *scute*, *lethal of scute* and *asence*) of the large achaete–scute complex (AS-C) and the *daughterless* (*da*) gene (Campuzano and Modolell 1992). These genes are of great importance for setting up the sensory organ pattern, i.e. where to form bristles in the adult fly, but are also involved in both the formation of the PNS and CNS. The neurogenic genes on the other hand, genes like *Notch* (*N*) or *Delta* (*Dl*), give the cells in the proneural cluster the potential to become neuroblasts, as part of the presumptive CNS.

The AS-C genes are expressed spatially restricted in clusters, the proneural clusters. In each of such clusters, which can be found in the imaginal disc cells and the abdominal histoblasts, single precursor cells, the sensory organ precursors (SOPs), get singled out during the third instar larval or the early pupal stages. They are in the process of differentiating to neural precursors.

The selection of a single cell to become the neural precursor involves a complex system of lateral inhibition (Simpson 1990). By passing on a signal to the neighbouring cells in the cluster, other cells are prevented from becoming competent to adopt the neural fate and will then give rise to epidermal cells.

The SOPs, also called the sensory mother cells (SMCs) undergo two divisions in order to eventually form a sensory organ each typically consisting of four cells, one neuron and a set of three support cells: a bristle cell (tricogen), a glial cell (thecogen), a socket forming cell (tormogen).

Since the adult cuticle displays a stereotyped reproducible pattern, an underlying mechanism must be present to produce that effect. A temporal sequence of events in the appearance of sensilla during development has been found.

Research carried out on the adult sensilla development has mainly concentrated on the notum, wing and the interommatidial bristles. It has been established that macrochaetes, in particular the large landmark bristles of the thorax and the recurved (chemosensory) bristles of the wing margin, form earlier than the microchaetes and the mechanosensory bristles of the wing margin (Hartenstein and Posakony 1989).

It has been proposed that the larger bristles need to form first in order to give spatial cues to the microchaetes, which will position and fit themselves evenly in between the macrochaetes. The precursors of the macrochaetes of the notum and wing start dividing at approximately the time of puparium formation. On the other hand, the microchaetes of the same area will only undergo their division between 9 h and 18 h after puparium formation. The precursors of the abdominal bristles are the last ones to divide between 24 h and 30 h after puparation. This is consistent with the development of the imaginal discs of the wing and notum compared with that of the histoblast nests, which will only start to differentiate 3 h after puparium formation.

1.6 Short outline of the following chapters

In the following chapters the analysis of the *smooth* mutant is described.

In chapter 2 the methods used are listed.

Chapter 3 contains further background on the initial discovery and isolation of the *smooth* (sm^3) mutant, including the description of the three different *smooth* phenotypes.

The subsequent chapter, chapter 4, describes the initial cloning carried out by Shrimpton, followed by the identification of the gene, further cloning, subcloning, sequencing and the description of the organisation of the gene carried out by myself.

Chapter 5 contains an analysis of the smooth protein, comparing it to homologous proteins and making some assumptions about its possible biochemical role or function. A discussion follows in the subsequent chapter 6.

Chapter 2

METHODS

2.1 Vectors: Fly strains, *Escherichia coli*, Bacteriophages and Plasmids

2.1.1 *Drosophila* stocks and strains

All lines were maintained in bottle and vial cultures on cornmeal–sucrose–agar medium at 25°C.

The wild-type strains used were the following: Harwich P-cytotype and Canton-S M-cytotype as described by Mackay (1985) and Samarkand M-cytotype (Lai and Mackay 1990).

The *Cy/sm*^{3} line was derived from a P-element-containing low abdominal bristle line, which resulted from an experiment selecting for abdominal bristle number with the help of transposable elements (Mackay 1985). For hybrid dysgenesis, the source of P-cytotype had been the Harwich strain and the source of M-cytotype had been the Canton-S strain (Mackay 1984; 1985). After a second chromosome extraction on the original low abdominal bristle line, all the P elements had been outcrossed, apart from two located at the cytological positions of 56E and 57B on the right arm of the second chromosome.

The *Cy/sm px* stock was obtained from the National Drosophila Species Center at Bowling Green. *sm*^{lab} (renamed *sm*^{2} by Lindsley and Zimm 1990) was derived from a Canberra outbred population and was a gift of R. Frankham (Frankham and Nurthen 1981).

2.1.2 Bacterial stocks

A number of different *E.coli* strains were required for experimental techniques, often depending on the cloning vector involved.

These *E.coli* strains were maintained as glycerol stocks, which were prepared as a mixture of 50% overnight culture of a single colony inoculation into L-broth grown overnight at 37°C and 50% of glycerol. The glycerol could then be stored long term at −20°C.

In order to obtain fresh single colonies from the glycerol stock, a small amount of it was streaked out on a Petri dish containing L-agar medium.

When plating cells were required, i.e. for infection of *E.coli* with bacteriophage λ, a single colony was inoculated into 10ml of T-broth containing 0.2% maltose and incubated with shaking at 37°C overnight. An aliquot, 0.1ml, was added to 10ml of T-broth and maltose and grown for a further 4–6 h. The culture was then centrifuged at 3,000 rpm for 10 min and resuspended in 5ml of 10 mM $MgCl_2$ to obtain a density of $OD_{600} \approx 2$.

The maltose which is present during the growth of the bacteria improves the efficiency of bacteriophage λ adsorption, since the maltose induces production of the λ receptor (lamB protein). Mg^{2+} ions play an important part in the maintenance and integrity of phage particles (Maniatis et al. 1982) and also aid phage adsorption.

2.1.3 Bacteriophage stocks

Bacteriophage stocks were stored as liquid lysates.

A single plaque was picked from a plate and placed into 0.2 - 0.5μl of PSB. After allowing the phage to elute into the medium, a dilution was being used to inoculate 100μl of plating cells.

After an incubation period of 15 min, allowing the phage to adsorb to the bacteria, the bacteria-phage-mix was transferred to a conical flask containing 10ml of L-broth plus 10 mM $MgCl_2$ and grown overnight at 37°C, shaking vigorously.

When lysis had been obtained, chloroform was added to lyse any remaining cells and shaking was continued for a further 15 min. The cell debris was then spun down and the supernatant was stored in a glass universal at 4°C after the addition of a few drops of chloroform.

In order to check the titre of the lysate, plating cells were inoculated with dilutions of the lysate and incubated at 37°C for 15 min. The inoculate was added to L-top agar, kept molten at 50°C, mixed and poured onto L-agar plates. These were then incubated at 37°C overnight.

2.1.4 Plasmids

Plasmids were stored long term with their host as glycerol stock at −20°C or as a stab at 4°C. Plasmid DNA preparations were either stored at −20°C or at 4°C.

2.2 Preparation of DNA

2.2.1 Preparation of genomic DNA from *Drosophila*

Approximately 500 flies (kept frozen at −70°C) were homogenised with several passes of the pestle in a glass homogeniser (Kontes Glass Co.) in the presence of 4ml of buffer containing 10 mM Tris-HCl pH 7.5, 10 mM NaCl, 10 mM EDTA, 1% SDS, 0.15 mM spermidine and 0.15 mM spermine, which gently lyses the nuclei.

An equal volume of a second buffer was added containing 100 mM Tris-HCl pH9, 30 mM EDTA, 2% SDS and 0.2 mg/ml of pronase-E (Sigma), the latter causes proteins to digest and to release the nucleic acid into aqueous solution.

After an incubation period of 1–2 h at 37°C, successive phenol (2×), phenol/chloroform (1×) and chloroform (1×) extractions were carried out in order to remove proteins.

The aqueous phase containing the DNA was ethanol precipitated twice by adding 0.7 M ammonium acetate and two volumes of absolute ethanol (cold), with the first centrifugation for 10 min at 3,000 rpm and the second one for 10 min at 5,000 rpm.

After the last precipitation, the pellet was washed in 70% ethanol and dried at room temperature for several hours.

The DNA pellet was finally redissolved in 1 μl TE for each original fly, i.e. 500μl.

2.2.2 Preparation of small-scale plasmid DNA (miniprep)

Small scale preparations of plasmid DNA (minipreps) were prepared by a modification of the boiling method originally described by Holmes and Quigley 1981.

This method relies on weakening and breaking of the *E.coli* cell wall by treatment with a combination of lysozyme, EDTA and heat. Cell lysis is induced by adding a non-ionic detergent, Triton X-100.

A single bacterial colony containing a plasmid was inoculated into 10ml of L-broth containing an antibiotic, if necessary. After an overnight incubation shaking at 37°C, 1.5ml of the culture was briefly spun down. The pellet was resuspended in 200μl of STET and 20μl of lysozyme (10 mg/ml). After boiling the suspension for 40 s at 100°C and centrifugation for 10 min, the flocculent pellet obtained was removed with a Gilson tip.

The supernatant was then precipitated in 0.3 M NaAcetate and 200μl of propan-2-ol, placed at −70°C for 15 min and spun down for 10 min. The pellet was rinsed with ether and dried under vacuum before resuspension in 50μl of TE.

2.2.3 Preparation of large-scale plasmid DNA (maxiprep)

For a large-scale preparation of plasmid DNA (maxiprep), the alkaline lysis method (a modification of Birnboim and Doly 1979) was combined with cesium chloride equilibrium density gradient centrifugation, to obtain a large amount of pure plasmid DNA.

This method is based on separation due to differences in conformation of plasmid (supercoiled) and *E.coli* DNA (linear and open circle). The fact that ethidium bromide binds in larger amount to linear DNA compared to supercoiled DNA is being exploited.

A 10ml overnight bacterial culture grown from a single colony was inoculated into 400ml of L-broth containing the appropriate antibiotic and vigorously shaken overnight at 37°C.

After centrifuging the bacteria in a Sorvall RC2-B for 10 min at 7,000 rpm, the supernatant was removed and the pellet was resuspended in 9ml of GTE. A 40 mg/ml solution of lysozyme was prepared in 1ml of GTE and added to the suspension. Following an incubation period of 10 min at room temperature, 80ml of freshly prepared NaOH (0.2 M) and SDS (1%) was mixed by swirling gently and chilled for 5 min on ice.

40ml of cold solution III was added, while still on ice. After 30 min, 10ml of H_2O was added before centrifuging at 8,000 rpm for 5 min. The supernatant was then sieved through a tea strainer and 0.6 volume of propan-2-ol was mixed in. After spinning for 5 min at 8,000 rpm, the supernatant was decanted and the pellet was resuspended in 11ml of TE, 240μl of 0.5 M EDTA and neutralised with 150μl of 2 M Tris Base (pH>7).

Eventually the volume was made up to 13.4ml with TE and the solution was added to 14.8 g of CsCl. After a further addition of 1.4ml ethidium bromide (10 mg/ml), the contents were transferred to an oakridge tube. The balancing of tubes is critical at this point.

Ultracentrifugation was carried out in a Beckman TY65 rotor at 35,000 rpm. After 72 h the lower band of supercoiled DNA, identified in the dark with a long-wave UV handheld lamp, was collected with a siliconised Pasteur pipette. The ethidium bromide bound to the plasmid DNA was extracted, either by passing the sample through a 4ml Dowex AG50W-X8 (Bio-rad) ion exchange column or by several extractions with n-butanol. After ethanol precipitation, the plasmid DNA was stored at 4°C in TE.

2.2.4 Bacteriophage λ DNA from minilysates

3.2ml of liquid lysate, which is a suspension of phage particles, was aliquoted into four Eppendorf tubes and 1μl of DNase I stock solution was added to each. During an incubation period of 20–30 min at room temperature, only the bacterial DNA will get degraded, since the phage DNA is protected by its protein capsid.

200μl of Tris-EDTA-SDS (0.3 M Tris-HCl pH 9.0, 0.15 M EDTA pH7.5 and 1.5% SDS) was mixed into the suspension, which was then incubated at 70°C for 15 min. After cooling at room temperature for 5 min, 135μl of 8 M potassium acetate solution were added to each tube. The samples were chilled on ice for 15 min.

The supernatant was removed after a short spin of 1–2 min in a microfuge and after an addition of 480μl of propan-2-ol, the tubes were incubated at room temperature for 5 min.

By spinning the samples for 2 min, a pellet was obtained, which was washed with 70% ethanol, dried under vacuum and resuspended in 50μl of TE each.

The contents of the four tubes were pooled and a single phenol extraction was carried out for deproteinisation purposes (removal of phage capsid). This was followed by a further ethanol precipitation and a final resuspension in 100μl of TE. The yield was checked by running aliquots against known amounts of DNA. The liquid minilysates are stored at −20°C.

2.2.5 Preparation of large-scale bacteriophage λ DNA

Plating cells were prepared for the required *E.coli* strain as described above and the liquid lysate was titred. Plating cells were inoculated into 400ml of L-broth containing 10 mM $MgCl_2$, so that the initial $OD_{600} \approx 0.1$. The culture was then shaken vigorously until the OD_{600}=0.4 . At this time point 2 × 10^8 pfu/ml of phage were added and and the shaking was continued. The growth of the cells was frequently monitored by taking samples and checking the OD_{600}. After 4–6 h, a peak (OD_{600}>2) was reached and the cultures started to lyse (OD_{600}>1.0).

In order to complete the lysis, 200μl of chloroform, 10 μg/ml DNase and 10 μg/ml RNase were added and the culture was shaken for a further 15 min. After a centrifugation period at 4,000 rpm for 10 min, 40 g/ml of NaCl were dissolved in the supernatant. Since phage particles are extremely small, 7% of crushed polyethylene glycol (PEG) 8000 was added, which in the presence of salt and water causes phage particles to precipitate.

This precipitation can be carried out for at least one hour or overnight at 4°C.

The phage was then pelleted by spinning at 8,000 rpm for 20 min and resuspended in 5ml of TMN. After two equal volume chloroform extractions, the phage suspension was made 41.5% CsCl in TMN and spun at 30,000 rpm overnight.

During this CsCl density gradient centrifugation, bacterial debris and unwanted cellular DNA can be separated from the λ particles. The phage band in a CsCl gradient at 1.45 to 1.5 g/cm^3 and can be seen as a bluish band.

The phage were collected from the oakridge tube with a siliconised Pasteur pipette and dialysed against 10 mM Tris pH 8, 5 mM NaCl, 1 mM EDTA to remove the CsCl. This was followed by three phenol extractions to digest the phage protein coat. Finally, the phage was precipitated by adding 0.7 M ammonium acetate and two volumes of ethanol.

2.3 Introduction of plasmid or bacteriophage λ DNA into *E.coli*

2.3.1 *In vitro* packaging of bacteriophage λ DNA

In vitro packaging of λ DNA is achieved by mixing extracts prepared from bacteria infected with a λ strain (BHB 2690 Eam) defective in one protein (i.e. E protein, required for assembly of prehead) and one strain (BHB 2688 Dam) defective in a second protein (i.e. protein D, involved in the insertion of λ DNA into the head precursor and maturation of the head, Hohn 1979).

An *in vitro* packaging kit containing the extracts was obtained from Amersham Intl. and the manufacturer's instructions were followed. The maximum yield obtained was between 2×10^8 to 2×10^9 plaque forming particles /μg.

2.3.2 Transformation of *E.coli* by plasmid DNA

Preparation of competent cells

Competent cells are required for transformation. These cells are 'competent', because their uptake of DNA is more efficient due to the presence of calcium ions. $CaCl_2$ affects only the DNA binding to the *E.coli* cell. The movement of DNA into the competent cells is achieved by heat-shocking.

The preparation of competent cells was carried out as described by Hanahan (1985) with some modifications.

Cells were plated out onto ψ agar (see Appendix) plates and grown overnight at 37°C. Initially, one colony was used to inoculate 10ml of ψ broth. After 2 h incubation at 37°C with moderate agitation, 5ml were subcultured in 95ml of ψ broth. Once the cells had reached an OD_{550} of 0.35 to 0.60, the culture was chilled on ice for 5 min before aliquoting the contents into corex tubes and chilling on ice for a further 10 min.

The cells were pelleted by centrifugation at 6,000 rpm for 5 min in a cold Sorvall. After discarding the supernatant, the pellet was resuspended in 10ml of TfbI (see Appendix).

After an incubation period of 15 in on ice, the cells were centrifuged at 6,000 rpm for 5 min. Each pellet was then resuspended in 1ml of TfbII (see Appendix). The suspension was aliquoted into Eppendorf tubes and quick frozen in liquid nitrogen before storing at −70°C.

Transformation

Differing amounts of plasmid DNA, ranging from 0.1 ng to 10 ng, were added to 30 to 100μl of competent cells slowly thawed on ice.

After incubation for 30 min on ice, the cells were heat-shocked by placing the tubes into a 37°C waterbath for 90 s. This was followed by 90 s of chilling on ice and the addition of 50 -100μl of ψ broth, before placing the tubes into a 37°C waterbath to incubate for a further 30–60 min.

If blue and white selection was required, 15μl of 10 mg/ml of IPTG (Isopropyl-β-D-thiogalactopyranoside, Sigma) and 40μl of X-gal (5-Bromo-4-chloro-3-indolyl-β-D-galactopyranoside, Boehringer Mannheim) 20 mg/ml was added. Plating was carried out on antibiotic containing medium which allowed the identification of plasmid containing colonies.

2.4 Manipulation of DNA

2.4.1 Digestion of DNA with restriction endonucleases

Restriction digests were carried out using varying amounts of DNA from 100 ng (e.g. minipreps) to 3 μg (e.g. genomic DNA or DNA for elution or genecleaning).

The units of enzyme per digest were adjusted appropriately varying from 4-20. Buffers were used as recommended by the supplier (Boehringer Mannheim).

Digests were carried out in a waterbath at 37°C for 1–12 h, with the exception of the enzyme '*Sma*I', which digests at 26°C. The enzymes were inactivated by incubation at 70°C for 5 min. Ficoll stop buffer (FSB) was added to the digests before loading them onto a gel.

2.4.2 Agarose gel electrophoresis

In order to analyse digested and undigested DNA (and also RNA, see below), horizontal gel electrophoresis (McDonnell, Simon and Studier 1977) was performed. The separation of fragments was carried out in an electric field, with the negatively charged DNA loaded at the cathode. Once a current is applied, the DNA migrates through the gel towards the positive electrode and the smaller DNA molecules migrate faster. A separation according to the size of fragment is taking place. The concentration of agarose had to be chosen appropriate for the size of DNA fragments (ranging from 0.5 to 25 kb) to be separated.

The gel was prepared by adding agarose to TBE buffer and melting it in the microwave. 1 μg/ml of ethidium bromide was added. After the gel was poured and set, it was immersed into TBE buffer containing also 1 μg/ml of ethidium bromide. The samples were then loaded with DNA markers (λ DNA either digested with *Pst*I or *Hind*III) usually on either side.

The gel was run until the samples had migrated sufficiently, as judged from the distance the dye had travelled. The DNA was visualised by placing the gel onto a UV illuminator.

2.4.3 Isolation of DNA fragments from an agarose gel

Genecleaning

A 'Geneclean Kit' was obtained from BIO 101 Inc. and the supplier's instructions were followed.

Genecleaning involves the excision of a band of DNA from a low melting point (LMP) TAE agarose gel. In the presence of NaI and 'glassmilk' a silica matrix is formed to which only DNA will bind. After several washing steps, DNA is eluted into TE free from RNA and protein contamination.

Elution of DNA fragments onto DEAE paper

The removal and purification of DNA from a gel by electrophoresis onto DEAE cellulose paper (Schleicher and Schuell) is an alternative to genecleaning.

A small piece of DEAE paper (e.g. 4 mm x 6 mm) was inserted into the gel, next to the band to be eluted. This operation was carried out under long-wave UV light, in order not to denature the fragment. The complete gel was then rotated by 90°, so that when electrophoresis was continued, the fragment migrated onto the paper.

Afer removing the paper and rinsing it twice in TE, 30–60μl of TE containing 1 M NaCl was added, in order to elute the fragment from the paper. The elution was carried out at 65°C for 30 min. If the eluted DNA was intended for use in a ligation or sequencing reaction, the high salt content had to be removed by spinning the elutant through a Sephadex G–50 column.

2.4.4 Southern blotting

An adaptation of the method devised by Southern (Southern 1975) was used to transfer DNA bands from agarose gels onto a nitrocellulose membrane. After the fragments were separated by gel electrophoresis, the gel was immersed in 0.2 M HCl for 20 min, which aids the cleaving of the larger fragments. The acid treatment could be omitted when the fragments were smaller than 5 kb.

The gel was then soaked for 30 min in 0.5 M NaOH, 1.5 M NaCl to denature the DNA, followed by a neutralisation step for 60 min in 1.5 M NaCl and 1 M Tris pH7.5.

The gel was placed on a support with a piece of Whatman filter no.17 in between the support and the gel. The filter paper served as a wick, with both ends dipped into 20× SSC. A piece of nitrocellulose membrane (Hybond N, Amersham Intl.), which had been cut to exactly the same size as the gel, was placed on top of the gel, followed by three sheets of Whatman filter no.3, presoaked in 2× SSC and a stack of 5–10 cm of paper towels with a glass plate and a weight (usually a heavy glass bottle) on top. All the filter paper and the paper towels were the same size as the gel.

The transfer was then allowed to proceed for a period between 3 h up to 12 h, before the membrane was rinsed in 2× SSC and either baked for 2 h at 80°C or UV cross-linked for 3–5 min to completely bind the DNA onto the membrane.

2.5 Preparation and Analysis of RNA

2.5.1 Collection of *Drosophila* eggs, larvae, pupae and adults

Collecting eggs

Eggs were collected on red wine concentrate or apple-juice medium (Wieschaus and Nüsslein-Volhard 1986) contained in petri dishes. The plates, with a slurry of yeast in the centre, were placed into the cage with the flies. After allowing the females to lay eggs for 12–20 h at 25°C, the plates were collected and the embryos were washed off with distilled water and transferred onto a nylon mesh in a little egg basket.

After several washes with distilled water the embryos were immersed in sodium hypochlorite for 3–5 min for dechorionation. Subsequently, the embryos were washed thoroughly in distilled water and quickly frozen on liquid nitrogen before storing them at -70°C until required for RNA preparation.

Collecting larvae

Petri dishes containing apple-juice or red-wine concentrate medium were placed into the cage as above. The plates were collected after leaving the flies to lay eggs for 12–20 h. The eggs were then transferred to larger petri dishes (diameter of 15 cm) containing plenty of live yeast. Those petri dishes were kept until the larvae had reached an age of 70–90 h. The third–instar larvae were rinsed off the medium with distilled water, washed thoroughly, immersed into liquid nitrogen and stored at -70°C.

Collecting pupae

The collection of embryos was carried out as above. After the embryos were transferred to the large petri dishes and had reached an approximately 70 h of age, the larvae, still on the apple-juice or red-wine medium were put into sandwich boxes containing 2 cm of fly food and a thick slurry of live yeast on top.

Buoyant pupae, aged between 140–160 h after egg laying, were collected by flooding the box with distilled water and collecting the floating pupae with the sieve. The pupae were immersed into liquid nitrogen and kept at -70°C.

Collecting adult flies

The adult flies were anaesthetised using either ether or carbon dioxide and then frozen at -70°C until they were used for either DNA or RNA extraction.

2.5.2 Preparation of total RNA

Total RNA was extracted by two different methods.

Quick phenol-chloroform extraction method

The following method was initially used for the preparation of RNA from SAM larvae and pupae. To approximately 5 g of larvae and pupae, 5–10ml of Savakus buffer (0.1 M Tris pH 7.5; 10 mM EDTA; 0.35 M NaCl; 7 M urea and 2% SDS) was added, followed by an equal volume of phenol-chloroform. The frozen larvae or pupae were then homogenised with a 'Polytron' tissue disrupter (Ystral type X1020), which has a cylindrical probe plus rotating cutters which cause a big vortex.

The samples were spun for 30 min at 8,000 rpm and equal volume chloroform extractions followed and were continued until no interface was visible any longer.

The aqueous top layer was then ethanol precipitated in 0.3 M NaAcetate and two volumes of ethanol and placed in a -20°C freezer for 1 h, before spinning at 8,000 rpm for 10 min.

The pellet was finally resuspended in TE containing 0.1% SDS. The RNA was stored as an ethanol suspension in 0.3 M NaAcetate and two volumes of ethanol.

When a RNA sample was required, aliquots of the suspension were spun down and resuspended in sdH_2O (ribonuclease-free).

The RNA concentration was measured at 260 nm, where an OD_{260} value of 1 equals 40 μg/ml.

RNA isolation by acid guanidium thiocyanate-phenol-chloroform extraction

This method was used for extraction of RNA from eggs, pupae, larvae and adult flies and is described by Chomczynski and Sacci 1987.

2.5.3 Extraction of mRNA

Two different kits were used to extract the mRNA from total RNA: the "RNA Purification Kit" from Pharmacia, which provides oligo(dT) columns and the "PolyATtract^{TM} mRNA Isolation System" from Promega, where the mRNA hybridises to biotinylated oligo(dT) and is then separated by streptavidin coupled to paramagnetic particles. The mRNA is eventually eluted in nuclease-free H_2O (provided by the supplier) from the solid phase.

The manufacturer's instructions were followed.

2.5.4 Northern blotting

RNA gel electrophoresis

RNA or mRNA was usually analysed on a 1.4% agarose gel.

In order to prepare 100ml of gel solution, 1.4 g of agarose was dissolved in 80ml of sdH_2O by microwaving. The gel was then cooled down to 60°C in a waterbath, before 2.0ml of NaPhospate buffer (NaPB) 0.5 M pH7 (final concentration of 10 mM) and 18ml of 38% formaldehyde was added. The gel was poured as soon as the above was mixed.

The RNA or mRNA sample to be loaded was made up to 10μl with sdH_2O, 2μl of 100 mM NaPB pH7, 7μl of 38% formaldehyde and 10μl of BRL grade formamide were added.

The sample was then heated to 55°C for 15 min before quenching on ice. 5μl of stop buffer was added.

The gel was immersed in 10 mM NaPB and run at 30–40 V overnight.

Blotting

The transfer was carried out similarly to the Southern transfer, but the soaking steps were omitted. The gel was therefore immmediately placed onto a support with Whatman paper no.17 in between. That paper served again as a wick and was soaked in 20× SSC.

A nitrocellulose membrane (Hybond N, Amersham) was placed on top of the gel, together with three Whatman no.3 filters soaked in 2× SSC and a stack of paper towels. Finally, a glass plate and a bottle were placed on top of everything else.

The blot was usually carried out overnight. The nitrocellulose membrane was then rinsed in 2× SSC, before baking it for 2h at 80°C or UV cross-linking for 3–5 min, in order to bind the RNA to the filter.

2.6 Polymerase chain reaction

The polymerase chain reaction (PCR) was predominantly employed when specific fragments of the cDNA were required as probes for screening genomic libraries and Southern blots, in order to work out the organisation of the gene. A list of primer sequences (see Appendix), mainly 18 to 20 mer oligonucleotides provided by Oswel DNA Service (Edinburgh University) were used for PCR and for sequencing reactions.

PCR amplification was carried out using TaqI polymerase (acquired either from Northumbria Biologicals Limited or Promega) at 25 U/ml. The 10× reaction buffer was supplied by the manufacturer (the same buffer applies to both manufacturers: 100 mM Tris-HCl, 50 mM potassium chloride, 15 mM magnesium chloride and 1% Triton X-100). Other components were added at the following concentrations: 17 μg/ml of bovine serum albumin, 10% DMSO, 33μM each of dGTP, dATP, dTTP and dCTP and 0.5 μM of each primer. The reaction was overlaid with approximately 100μl of paraffin oil (BDH).

Amplifications were performed in a PCR machine (Techne PHC-2) which was set for 25 s at 94°C for the denaturing reaction, 35 s at 50°C for the annealing reaction and allowing 2.5 min at 68°C for the extension step.

After 25 cycles, a chloroform extraction was carried out, in order to remove the paraffin oil. Subsequently, the PCR product was genecleaned, removing the unincorporated dNTPs and primers and concentrating the DNA. An aliquot was checked on a gel.

2.7 Synthesis of radioactively labelled probes

2.7.1 Nick-translation

Nick-translation of fragments of DNA (Rigby et al. 1977) or complete plasmids is one method of labelling DNA. It exploits the fact that *E.coli* DNA polymerase I is capable of both polymerase (i.e. adding nucleotide residues to the 3' terminus of a nick in the 5' to 3' direction) and exonuclease activity (i.e. removal of nucleotides from the 5' side of the nick proceeding towards the 3' direction).

Therefore, if a nick is produced in double-stranded DNA by the addition of a small amount of DNaseI to the reaction, DNA polymerase I will synthesise a completely new strand in the presence of nucleotide, as the existing strand will be degraded as the DNA polymerase I proceeds in the 5'to 3' direction. If a radioactively labelled nucleotide is supplied, it will be incorporated and the fragment is labelled.

Nick-translation was frequently used to label whole plasmids which had a fragment of interest inserted.

In a reaction with a final volume of 100μl, 100 ng of DNA was labelled by the addition of 10μl of nick-translation-buffer (NTB), 5μl of 0.5 M dNTPs -dCTP, 5μl of 3000 Ci/mmol [α–^{32}P]dCTP (50μCi from Amersham Intl.), 1μl DNaseI (diluted

10,000 fold from a 1 mg/ml stock solution), 1μl of 0.1 M dithiothreitol (DTT), 1μl of 0.4% bovine serum albumin (BSA) and 2μl *E.coli* DNA polymerase I (5 to 15 U).

Incubation was carried out for 2–4 h in a 14°C waterbath. The incorporation of label into the DNA was monitored, either by carrying out a Cerenkov count or by simply checking it in front of the Geiger counter. This was done by baking 1μl of DNA onto a small piece of nitrocellulose membrane. The incorporation was measured before and after dissolving excess free nucleotides in 5% trichloroethanoic acid.

In order to remove any unincorporated dNTPs, the sample was spun through a Sephadex G-50 column for 90 s at 1,500 rpm. Finally, the reaction mix was denatured in a third of the volume of 1 M NaOH and neutralised with an equal volume of 1 M Tris pH 7.5, before adding it to the prehybridisation solution and the filter (see below).

2.7.2 Random primer labelling

Random primer- or oligolabelling was usually carried out on a small linear fragment of DNA which was purified, either by elution from a DEAE membrane, by genecleaning or by labelling a band of DNA after melting in a block of LMP agarose.

Random primers (a mixture of hexanucleotides) were added to denatured DNA. In the presence of dNTPs and radioactive label, the Klenow fragment was used to synthesise a new complementary strand, incorporation the label. An 'Oligolabelling Kit' was obtained from Pharmacia and the manufacturer's instructions were followed, both for purified DNA and DNA in agarose.

The incorporation of label achieved was usually greater (>80%) than when nick-translation was carried out. The Sephadex step was omitted and the reaction was denatured and neutralised as described above.

2.7.3 Ribolabelling: SP6/T7 transcription

This method of labelling involves the synthesis of a radioactive RNA probe.

A fragment of DNA which should serve as a probe, has to be cloned downstream of a bacteriophage promoter (either SP6 or T7) in an appropriate transcription vector (e.g. pGEM2 from Promega). The plasmid is linearised by cleaving with a restriction enzyme which only cuts 3' of the inserted fragment on the opposite side of the promoter. SP6 or T7 RNA polymerase is added, which will consequently transcribe DNA sequences downstream of the SP6 or T7 promoters, respectively.

In order to purify the RNA produced, the DNA template is removed by the addition of DNaseI.

The radioactive label required in this case is a ribonucleotide triphosphate (Amersham Intl.). All ingredients (apart from the radioactive label) were included in the 'SP6/T7 Transcription kit for *in vitro* transcription of DNA' from Boehringer Mannheim.

2.7.4 Hybridisation

The baked membrane filter was placed into a polythene bag (Jencons Scientific). Depending on the size of the filter, 10–20ml of prehybridisation solution (0.5 NaPB pH7.2; 7% SDS and 1 mM EDTA) prewarmed to 65°C was added. The bag was sealed and placed into a 65°C incubator for 5 min to 12 h.

The ^{32}P-labelled, denatured DNA probe was then added to the prehybridisation solution in the bag and mixed, before carefully sealing the bag. Incubation was carried out overnight at 65°C.

Washing of filters

After an overnight incubation, the hybridisation solution was poured into a polythene universal tube, in order to be able to reuse it.

The filter was then washed in 1 mM EDTA; 40 mM NaPB pH7.2 and 1% SDS to remove any unbound and weakly binding probe. Three washes were carried out for 30 min each, before the filter was sealed in Saran wrap. The membrane was placed in a autoradiographic cassette, overlaid with a sheet of X-ray film (RP1 CURIX, AGFA) and stored at -70°C until the exposure was thought to be appropriate for the film to be developed.

The X-ray films were developed in an automatic X-ray film processor (X-ograph X1).

Reusing of a probe

The probe saved in the polythene tube was boiled for 20 min, left to cool down to approximately 65°C and then transferred to the filter, which had previously been prehybridising. The prehybridisation solution was discarded in this case. Probes were reused several times, especially when dealing with Southern blots of plasmid DNA.

Removal of probe DNA from membranes

The probe DNA could be removed from the nylon filter by boiling it for 3 min in 0.1% SDS. The filter was reused after the solution cooled down to room temperature.

2.8 Construction of genomic library

2.8.1 Partial digests, ligation and packaging of genomic DNA

A genomic library consists of a large number of recombinant clones which should be representative for the genome under investigation. In order to find out how many plaques (N) are required for creating a representative library for *Drosophila*, the following formula is applied:

$$N = \ln \frac{(1-P)}{(1-f)}$$

(Clarke and Carbon 1976).

P is the probability that a given gene is represented in the library, usually P= 0.99, and f is the size in bp of the desired fragments, as a fraction of the genome size. The approximate size of the partially digested genomic DNA is 1.5 × 10^4 bp and the size of the *Drosophila* genome is 1.5 × 10^8 bp.

$$N \ = \ln \frac{(1-0.99)}{1-(1 \times 10^4)} = \ 46,026$$

Partial digests of genomic *Drosophila* DNA were carried out, using the *Sau*3AI restriction endonuclease, which recognises a 4 bp site (GATC). Genomic DNA was digested for variable lengths of time with *Sau*3AI and a small aliquot of these digests were run on a 0.4% gel alongside λ *Hind*III marker. Digests which produced most DNA fragments in the range of 10–20 kb were concentrated by ethanol precipitation and consequently used to ligate into EMBL4.

The EMBL4 vector was prepared as described earlier (see large-scale bacteriophage preparation), digested with *Bam*HI, which produces a left arm of 19.4 kb, a central stuffer fragment of 13.7 kb and a right arm of 8.9 kb, and was ethanol precipitated.

The generated genomic *Sau*3AI fragments were then ligated into the left and the right arm of EMBL4 by the addition of 1 μg of EMBL4; 0.5 μg of genomic DNA (aiming at a concentration ratio of 2:1 between EMBL4 and genomic fragments); 1μl of 10 x ligation buffer an 1μl of T4 ligase (1 Weiss Unit) in a total volume of 10μl. A control ligation, which consequently served as a packaging control, was also set up. It consists of a self-ligation of EMBL4. The ligation reactions were incubated at room temperature for 3 h or at 14°C overnight.

An aliquot of the ligation mix was then checked on a gel by comparing it to unligated partially digested DNA. If the ligation was successful, *in vitro* packaging followed by using the Amersham "*In vitro* packaging kit", as described earlier. 2μl of the ligated DNA were utilised for packaging which took place at room temperature for 2 h. Another control was set up which consisted of packaging 1 μg of undigested EMBL4.

Phage suspension buffer (200μl) and one drop of chloroform were added to dilute the packaging reaction. The titre of the packaging mix was obtained by preparing dilutions and plating them out using a Q358 (Karn *et al.* 1979) strain and a P2 lysogen Q359 (Karn *et al.* 1979) strain. The efficiency of ligation and packaging could be calculated from comparing the number of plaques obtained from the controls and the library packagings.

Once the titre was established, an appropriate volume of the packaging mix was then plated out on 10 large plates (diameter of 15 cm) using the Q359 *E.coli* strain and incubated overnight at 37°C. The resulting plaques were directly screened by hybridisation.

2.8.2 Plaque hybridisation

The bacterial plaques were transferred to nitrocellulose filters. The plates were chilled at 4°C for a few hours to allow the top agarose to harden. Circular Hybond (Amersham) or nitrocellulose membranes (Schleicher and Schuell) were cut to the appropriate size and labelled, before carefully placing them onto the plaques with the help of a pair of forceps. The membranes were kept in that position for approximately two minutes, while producing an asymmetrical pattern on the membrane and the bottom of the petri dish by stabbing through the membrane into the agarose with a needle. The filter was then carefully peeled off, without removing the top agarose from the plate and placed for 2 min in a denaturing solution (1.5 M NaCl, 0.5 M NaOH) followed by 2 min in neutralising solution (1.5 M NaCl, 1M Tris-HCl pH8) and for at least another 2 min in 2× SSC. The filters were then transferred onto paper towels, dried, and baked at 80°C for 2 h.

Hybridisation to radioactive probes

Hybridisation to a nick-translated or oligolabelled probe was carried out as described above.

Washing filters

Washes were carried out as described earlier. The filters were then sealed in Saran wrap. Orientation positions were marked by either drawing a pattern with radioactive ink on a small white paper sticker or placing a Glogos autoradiogram marker (Stratagene) onto the Saran wrap. An X-ray film "RP1 curix" (AGFA) was then placed on top.

2.8.3 Identification and purification of positive bacteriophage clones

The autoradiograph, filters and the plates were properly aligned and the positions of positive plaques were marked on the base of the agar plates. If the positive could not precisely be located to one single plaque, the wide end of a Pasteur pipette was used to spear into the agar and the agar plug was suspended in 500μl of PSB. A drop of chloroform was added. After vortexing and resuspending the phage for 1–2 h several dilutions were prepared. Q358 plating cells were inoculated with those dilutions, plated on small agar plates and incubated overnight at 37°C.

A second screening was carried out by taking lifts of the plates with a low density, in order to be able to pick single, separate plaques. The filters were treated as above and hybridised to the probe used earlier.

The filters were exposed to an X-ray film for 4–16 h. Positive plaques, if found singly, were picked and resuspended in 200μl of PSB with the addition of one drop of chloroform. If no single plaque was available, a further screening had to be carried out.

Minilysates were set up by aliquoting 20μl and 100μl of the positive resuspended phage into 10ml of L-broth containing 10 mM $MgCl_2$. The cultures were grown overnight at 37°C, shaking hard. The preparation of DNA from lambda minilysates has been described above under 2.2.4. .

2.9 cDNA library

An amplified cDNA library of 12–24 h embryos was obtained. The library was prepared by Brown and Kafatos (1988). It is a plasmid library and the plasmid, pNB40, was constructed by N. Brown (Brown and Kafatos 1988).

2.9.1 Transformation of cDNA

A control transformation was carried out, in order to check the efficiency of the XL1-Blue competent cells. Transformation was carried out by adding 1μl of cDNA (5 ng/μl) to 100μl of XL1-Blue cells (efficiency of 1.2 x 10^7 cfu/μg). After an incubation period of 30 min on ice, the cells were heat-shocked at 42°C for 40 s and then chilled on ice for 90 s.

0.9ml of SOC medium (see Appendix) was added before placing the transformation mixture in a 37°C shaker (at 225 rpm) for 1 h. The transformation mix was then plated on large plates (diameter of 15 cm) containing 2 x TY agar with 100μg/ml of ampicillin. Incubation was carried out overnight at 37°C. In total, six transformations were prepared. Approximately 20,000 colonies were obtained on each plate.

Taking colony lifts and preparation of replica filter

A nitrocellulose filter was placed onto a library plate for 2 min. The filter was marked by piercing holes through it with a 20-G needle. A new nitrocellulose filter (the 'replica' filter) was placed on a L-amp plate. The other filter was removed from the library plate and carefully placed on top of the replica filter.

The replica filter was subsequently marked at exactly the same position as the original filter, by piercing holes through the marked positions. This "sandwich" was then incubated at 37°C. After 4 h, the filters were carefully peeled apart and then placed with the colony side facing upwards onto fresh L-amp plates containing chloramphenicol at 50 μg/ml in the top agarose. This chloramphenicol amplification is carried out in order to improve the signal produced during hybridisation. The plates were incubated overnight at 37°C. The original library plates were incubated at 37°C for 4 h and then transferred to 4°C.

Preparation of filters for hybridisation

The filters were denatured and neutralised by placing them in the following solutions for 2 min each:

- 0.5 M NaOH/1.5 M NaCl
- 1 M Tris-HCl pH 7.5/1.5 M NaCl
- 2× SSC + 0.1% SDS
- 2× SSC
- 2× SSC

The filters were then blotted and baked for 2 h at 80°C.

Prehybridisation was carried out for 20 min at 63°C in the hybridisation mix, as described above.

An oligolabelled 1.1 kb genomic fragment was added as the probe and incubation was continued overnight at 63°C.

Washes of filters were also carried out, as described above.

2.9.2 Purification of positive colonies

The "positive" colonies were picked with a sterile loop and inoculated into 1ml of L-amp. Dilutions of the suspension were prepared (e.g. 1μl in 800μl) and 100–150μl of this dilutions was spread onto a small L-amp plate. The colonies were grown overnight at 37°C. Lifts were taken and hybridisations were carried out, as described above, until a single positive could be purified. Minipreps and subsequently a maxiprep were prepared from the positive.

2.10 DNA sequencing

2.10.1 Preparation of helper phage

For the production of single-stranded DNA, a helper phage (e.g. R408, a derivative of M13 phage) is required which is used to infect a plasmid (e.g. pTZ19R/18R, Pharmacia) containing the intergenic region of filamentous phage f1. When infected, the plasmid enters the f1 replication mode and produces single-stranded DNA which is packaged and secreted from the cell as phage particles.

To optimise the yield of single-stranded DNA, the origin of replication of the R408 helper phage (Russel, Kidd and Kelley 1986) was partially inactivated, resulting in the production of more single-stranded DNA particles than R408 DNA.

Preparation of R408 helper phage A colony of NM522 or XL1-Blue (Stratagene) previously selected on tetracycline containing medium, was grown at 350 rpm for 4–6 h. A 500ml conical flask with 50ml of 2× TY broth was inoculated with 100μl of the NM522 or XL1-Blue cells and grown to A_{660}=0.1 at 37°C. 20μl of R408 (10^8 pfu/ml) were added and growth was continued at 37°C overnight.

The culture was spun at 3,000 rpm for 15 min. The supernatant was collected and the spinning was repeated. To the supernatant 500μl of chloroform was added. The helper phage was stored at 4°C. The helper phage was titred before use.

2.10.2 Preparation of single-stranded DNA

Plasmid DNA (e.g. pTZ19R/18R) containing cloned inserts were transformed and plated out. Single colonies were picked and inoculated into 10ml of 2× TY broth containing ampicillin (100 μg/ml). The culture was grown overnight. 10μl of it was used to inoculate 3ml of 2× TY broth in a 50ml conical flask. The inoculate was grown for 1 h at 37°C, shaking vigorously, before adding 10–20μl of the R408 helper phage stock (8×10^{11} pfu/ml). Shaking was continued overnight.

Aliquots of 1.5ml of the culture were spun for 10 min at 4°C. To 1ml of the supernatant, 200μl of PEG/NaCl solution (20% PEG; 2.5 M NaCl) was added. The mixture was allowed to stand at room temperature for 20–30 min, before spinning for 10 min in a microcentrifuge. The supernatant was discarded and the pellet was resuspended in 100μl of T.E.. A phenol extraction was followed by an ethanol precipitation and the pellet was finally resuspended in 20μl of T.E.. In order to check the yield, 5μl was loaded onto a gel.

2.10.3 Double-stranded nested deletions for sequencing

The introduction of double-stranded unidirectional nested deletions can facilitate the sequencing of a large DNA fragment inserted into a plasmid. After linearisation of the plasmid, creating either blunt or 5' overhanging ends, digestion with exonuclease III for varying length of time will progressively unidirectionally remove nucleotides from one strand of the target DNA. Subsequent SI nuclease treatment will digest any single-stranded DNA of the other strand, which had not been attacked by exonuclease III, i.e. the other strand. The final step in the procedure is the re-ligation of the two ends resulting in a series of plasmids containing DNA inserts of different length.

A single primer, located in the plasmid close to the site of re-ligation, should now be sufficient for sequencing the complete insert.

The reactions were carried out according to the instructions supplied by the manufacturer of the 'Double-stranded nested deletion kit' (Pharmacia).

2.10.4 Sequencing of single- and double-stranded DNA

Sequencing of single and double-stranded DNA was carried out using the dideoxy method (Sanger et al. 1977). The primers used for sequencing were either - M13 primer or internal primers of specified sequence, which were synthesised by the Oswel DNA Service, Department of Chemistry, University of Edinburgh, and were purified by high-performance liquid chromatography.

Single-stranded sequencing was carried out by using the Sequenase-2 kit (United States Biochemicals) with [^{35}S]dATP (Amersham; >1000 Ci/mmol) according to the manufacturer's instructions.

Double-stranded sequencing on minipreps (without removal of RNA), maxipreps, genecleaned or eluted fragments of DNA was performed first of all denaturing 3–5 μg of DNA in 0.2 M EDTA and 0.2 M NaOH, as described by USB. The standard annealing and termination reactions were modified by the addition of dimethyl sulfoxide (DMSO) (Winship 1989) to a final concentration of 10% and 7.5%, respectively.

After boiling the annealing mix for 3 min, the samples were immediately chilled on ice for 3 min. The labelling reaction was performed at room temperature for 5 min and the termination reaction for 10 min at 37°C.

The samples were loaded onto a 8% polyacrylamide denaturing gel. A wedged gel was frequently used, which allows more nucleotides to be read from a single loading.

Most of the sequencing was carried out using an electrophoresis apparatus called "Base runner" (Internationals Biotechnologies, Inc.). Previously a single Perspex stand and two pieces of window glass had served the purpose. According to the manufacturer's recommendations, electrophoresis on the "Base runner" was carried out at a constant power setting of approximately 45 W, which supposedly held the temperature of the gel at ca. 55°C, a temperature sufficiently high to keep the DNA denatured and to reduce artifacts resulting from secondary structure.

The gels were routinely run for 4–6 h, extending the range of sequence readable to 200–300 nucleotides in one sequencing run. The gel was subsequently fixed in 10% methanol (BDH) and 10% glacial acetic acid (BDH) for 20–30 min, before drying it for 1–2 h at 78°C. A wedged gel was run between 4–8 h and dried for at least 2.5 h. Autoradiography using an "AGFA curix RP1" X-ray film was carried out, leaving the film to expose for 12–16 h.

Sequence analysis was performed on a VAX computer, using the University of Wisconsin GCG programs (Devereux *et al.* 1984).

2.11 *In situ* hybridisation to salivary gland chromosomes

The salivary glands of third-instar Samarkand larvae were dissected and *in situ* hybridisation was performed according to Leigh Brown and Moss (1987). This work was kindly carried out by Sarah Ross.

As the probe, the complete cDNA (2μg) in pNB40 was nick-translated according to Leigh Brown and Moss (1987). The polytene chromosomes were examined under oil emersion at a 100× magnification with the Olympus BH2 microscope.

Colour photographs were taken using the Olympus photographic system (model PM-10AD) with the automatic exposure control unit (PM-CBAD) at 40× and 100× magnification.

2.12 *In vitro* transcription and translation and SDS gel analysis

2.12.1 Transcription and translation

An *in vitro* transcription and translation kit was obtained from Promega and the instructions provided by the Promega "Protocols and Applications guide" were closely followed.

The template DNA (5 μg) for the *in vitro* transcription was the cDNA fragment, subcloned into pTZ19R and linearised with *Eco*RI. After a phenol/chloroform extraction and an ethanol precipitation, the DNA was resuspended in 1μg/μl, 4 μg of which was utilised in the T7 RNA polymerase transcription reaction according to Promega's recommendations. The transcription product was subsequently phenol/chloroform extracted and resuspended in 10μl of H_2O.

For *in vitro* translation, the wheat germ system (Promega) was utilised and [^{35}S] methionine (Amersham Intl.) at a final concentration of of 500 μCi/ml was incorporated as the radiolabelled amino acid.

The positive control for the translation reaction was the *Brome Mosaic Virus* (BMV) RNA which was supplied by the manufacturer.

Aliquots of translation products were stored at -70°C, if not used immediately for electrophoresis.

2.12.2 SDS gel analysis

Discontinuous gel electrophoresis under denaturing conditions in the presence of 0.1% SDS, a method for electrophoresis as described by Laemmli (1970), was carried out. Vertical slab minigels with a 12% SDS polyacrylamide separating gel and a 5% stacking gel were prepared for the Mighty Small II SE250 Hoefer Scientific gel electrophoresis apparatus.

To 5μl of the wheat germ extract translation product, 20μl of 2× loading buffer was added. The buffer contains β-mercaptoethanol, the disulfide reducing agent. In order to prevent any protease activity and to denature the proteins, the sample was heated at 100°C for 5 min after the addition of the loading buffer. 10μl of this mixture was subsequently loaded onto the gel alongside the BMV control translation and parallel to the ^{14}C-labelled Rainbow^{TM} protein molecular weight markers (Amersham Intl.) ranging from 14,300–200,000 M.W..

A constant current of 10 mA was applied to the samples whilst migrating through the stacking gel. When the samples entered the separating gel, the current was increased to 20–30 mA.

Once the bromophenol blue dye front had run off the bottom of the gel, the gel was removed and stained in 30% methanol, 10% glacial acetic acid containing 0.1% Coomassie Blue for 30 min. After a series of destaining in 30% methanol and 10% glacial acetic acid, the gel was layered on a sheet of Whatman 3 MM paper and dried for 1 h at 78°C. Autoradiography using an "AGFA curix RP1" X-ray film was carried out, leaving the film to expose for 12–16 h.

Chapter 3

The smooth phenotypes

3.1 Background I - a *smooth* mutation induced by hybrid dysgenesis

As one of the outcomes of an investigation into transposable element-induced response to artificial selection for number of abdominal bristles in *Drosophila melanogaster* carried out by Mackay (1985), a semi-lethal allele was identified on the second chromosome extracted from the low dysgenic selection lines.

This semi-lethal allele was characterised as being recessive, producing sterile females not maintainable as homozygotes. It was found independently at least twice in these lines.

Further investigations involved genetic mapping in order to locate the allele at a certain locus, since a number of P elements were still present in the selection line.

Heterozygous flies, *Cy* over mutant allele, were crossed to a strain of flies with *Cy* over seven mutations which were spread along the second chromosome. Analysing the result of these crosses suggested the mutation, caused by the P element, lay between *c* (curved at 2:75) and *px* (plexus at 2:100) at approximately 2:80. After consulting Lindsley and Grell (1968), the *smooth* locus, at 2:91.5 was considered as a good candidate for the gene (analysis carried out by Professor A. Robertson cited by McKay 1985).

A complementation test was carried out using a *Cy/sm px* stock (obtained from a stock centre) and the new allele. They failed to complement, indicating that a new allele at the *smooth* locus had been discovered.

In order to confirm that a P element insertion at 56E was indeed the cause of the *smooth* mutation, further genetic mapping experiments were required. Dr A. Shrimpton carried out more precise genetic mapping, using a multiple marked chromosome $L\ c\ nw^B\ Pu$ (L at 2:72, c at 2:75, nw^B at 2:83, Pu at 2:97) where the markers were more closely linked to the *smooth* locus than before.

The usage of dominant flanking markers allowed direct determination of F_2 males with a second chromosome which had undergone a cross-over in the interval between *nw* and *Pu*. These males were then crossed to *Cy/sm* females, so that their *smooth* phenotypes could be scored.

Out of 50 recombinants (30 between nw^B and *sm*; 20 between *sm* and *Pu*), 28 had a *smooth* phenotype which was also associated with a P element insertion at 56E, as shown by *in situ* hybridisation of P elements performed on each of the recombinants. The remaining 22 recombinants exhibiting a wild-type (sm^+) phenotype with regards to *smooth*, did not show a P element homology at 56E. The results obtained indicated that the P element was located within one map unit of the *smooth* locus (at a 95% confidence limit).

3.2 Background II - locating the P element

In order to determine the cytological location of the P element causing the mutation, 14 extracted second chromosome lines from the low dysgenic selection lines were isolated. These were scored for their bristle phenotype, i.e. whether they were *smooth* or not.

In situ hybridisations, using the P element as the probe, carried out by Dr A. Shrimpton identified nine lines also exhibiting a *smooth* phenotype. All of these lines showed a P element homology at the cytological location of 56E. Since the 56E cytological position could correspond, or at least approximate, to the 2:91.5 genetic position, it was considered to be the correct position where the P element could have been responsible for the mutation.

The other five low selection line extracted chromosomes did not produce a *smooth* phenotype and *in situ* hybridisations showed no P element insertion at position 56E.

The results of the above described experiments give a clear indication about the site of mutation. A P element insertion at the site of mutation is the most likely cause of that mutation.

3.3 The *smooth* alleles and phenotypes

1) The first *smooth* allele

Bridges and Brehme (1944) came across a spontaneous mutation in a *px pd* stock. The mutation was labelled *smooth* (*sm*) and located at 2-91.5. Their description of the *sm* phenotype can be summarised as follows:

- ventral surface of the abdomen partially denuded of bristles and shrunken
- wings usually warped and semi-erect
- acrostichal hairs disarranged
- tendency for erect postscuttellars
- male genitalia often disturbed
- anal protuberances of female bent down
- viability 30% of wild-type
- both sexes entirely sterile

The *sm px* chromosome was obtained as a *Cy/sm px Drosophila melanogaster* fly stock and now carries a lethal. The above described phenotype (*sm/sm*) was therefore never observed.

2) The second *smooth* allele, sm^{lab} or sm^2

Frankham and Nurthen (1981) detected a rare allele of large effect in a sample of a Canberra outbred population. Complementation tests showed that they had discovered a new allele at the 2-91.5 locus and it was called sm^{lab}, later renamed to sm^2.

The sm^{lab} phenotype:

- abdominal bristle number reduced in homozygotes by more than 50%
- altered pattern of abdominal bristles
- reversed sexual dimorphism for abdominal bristles (see below)
- normal fertility

The abdominal sternite bristle character is normally sexually dimorphic with the females having a larger number of bristles than the males. In the case of a reversed sexual dimorphism, the males show a larger number of bristles than the females.

3) The new *smooth* allele

In an experiment on transposable element-induced response to artificial selection in *Drosophila melanogaster* (Mackay 1985), a mutation was identified in low selection lines which was responsible for an extreme bristle phenotype, causing an almost complete lack of abdominal bristles in some flies.

This mutation was subsequently shown to be caused by a full-length P element insertion at cytological location 56E on the second chromosome (IIR). Genetic mapping experiments were carried out (A. Robertson, unpublished results) and since the mutant allele failed to complement *smooth* which provided to be a suitable candidate at 2-91.5 according to Lindsley and Grell (1968), it was clear that another allele of the *smooth* gene was discovered. It is being referred to as sm^3 or sm^P. A null mutant was not obtained.

Phenotype of sm^3:

- abdominal bristles almost completely absent
- allele is largely recessive
- homozygous sterile
- poor viability

Other features and pleiotropic effects observed at varying degrees are: acrostichal hairs lost, microchaetae from around the eye lost, aristae reduced and macrochaetae (including scutellars) weak and thinned, but the sternopleural bristles remained unaffected.

Description of the photograph of the abdomens

On the photograph (figure 1) three different *Drosophila* abdomens (a, b, c) are depicted, where a) shows the abdomen of a wild-type (Samarkand) female. The little rectangular plates along the midline of the ventral abdominal surface, the sternites, are clearly visible, in particular sternites three, four, five and six. The second sternite is partially covered by legs on that picture and the last sternite (number seven) is slightly out of focus.

Figure 1b) shows the abdomen of a female fly homozygous for the second *smooth* allele sm^2/sm^2 derived from the Canberra population. As in a), the clearly visible sternites are number three, four, five and six, with the second sternite again being covered by the fly's legs and the seventh sternite is difficult to identify due to the lack of contrast. Compared to the wild-type abdomen, it shows that the number of bristles is largely reduced and the bristles are shorter. In particular the third and fourth sternite show a notable absence of large chaetes.

The third part of figurel, 1c), shows the abdomen of a male *Drosophila* homozygous for the third *smooth* allele, sm^3/sm^3. Apart from a few black stubs, the shortened bristle remains, no real indication of the sternites is apparent. A complete null mutant, an abdomen without any bristles, has not been observed. The phenotype described for figure 1c) of the photograph can be obtained, either when *Cy/sm* (where the *sm* mutation is the original Bridges and Brehme (1944) mutation) flies are crossed to *Cy/sm*^{3} flies, or when *Cy/sm*^{3} are crossed to *Cy/sm*^{3}. Nevertheless, progeny with this phenotype are rarely obtained, since these *smooth* alleles are lethal to semi-lethal.

As already mentioned above, a cross between *Cy/sm* and *Cy/sm* only resulted in lethals, since the *Cy/sm px* stock now carries a lethal. But according to Lindsley and Grell (1968), the description of the abdominal phenotype seems to closely resemble that of photograph 1c).

When homozygous *sm*^{2}/*sm*^{2} flies are crossed to either of *Cy/sm* or *Cy/sm*^{3}, the non-curly flies survive and show *sm*^{2}-like phenotype, but never a phenotype as described for 1c).

Photograph: The ventral view of the abdomens of three different adult flies.

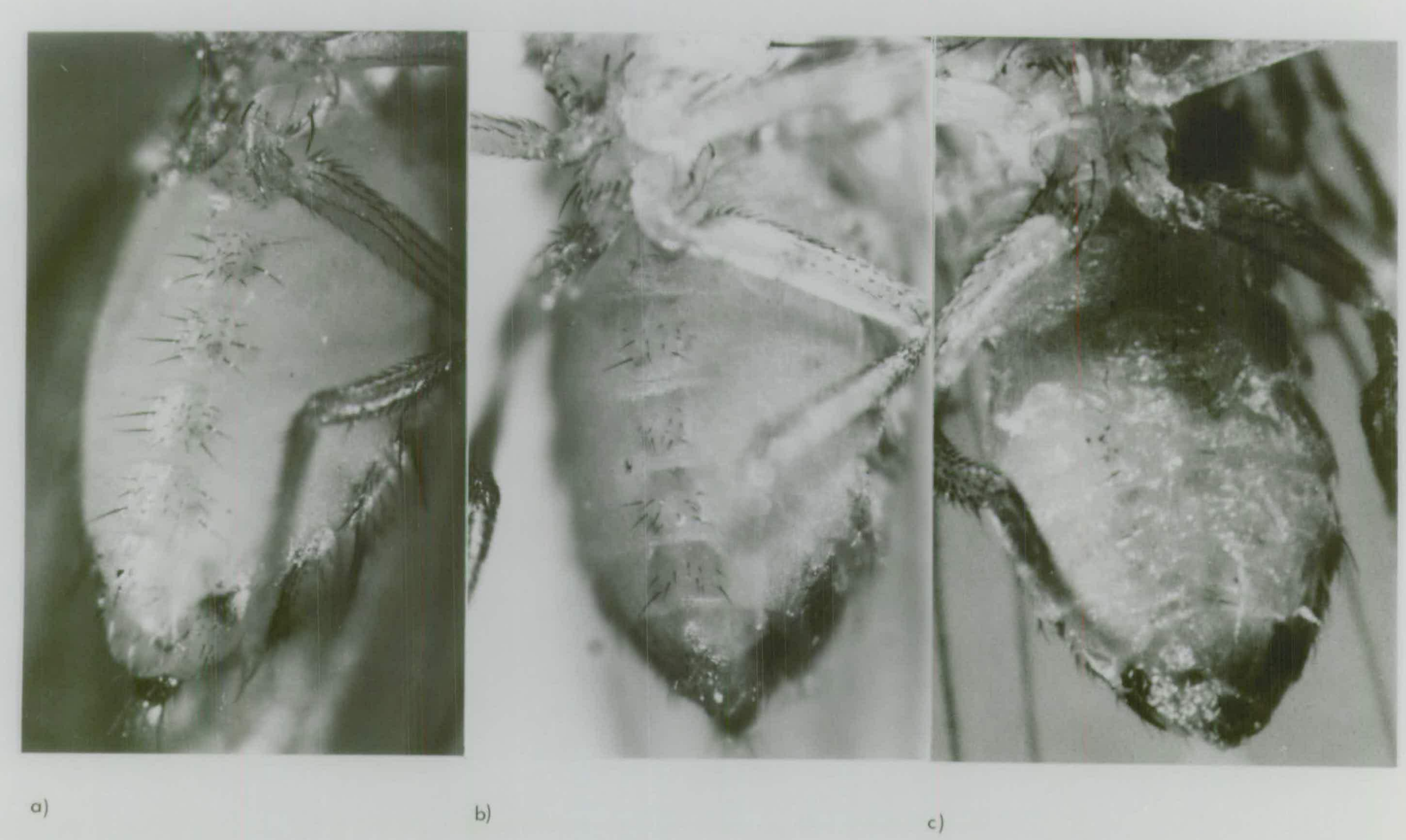
For detailed description see text above.

a) wild-type (Samarkand) female

b) sm^2/sm^2 *female*

c) sm^3/sm^3 male

Opposite page 75


a)
b)
c)

Chapter 4

Molecular analysis of the smooth locus

4.1 P element homology mapping

4.1.1 Background

This section summarises work carried out by Dr A. Shrimpton.

In order to facilitate the molecular analysis of the *smooth* locus, a *Drosophila* stock had to be generated which retained as few P elements as possible, apart from the insertion at 56E (sm^3). This was accomplished by taking a suitable recombinant male from an earlier experiment, $Cy/L\ nw^B\ sm^3\ Pu^+$, which carried the least P elements (according to *in situ* results), crossing it to *Cy/Pu* females before backcrossing Cy^+ female progeny to *Cy*. Individual $Cy/L\ nw^B\ Pu$ progeny were crossed to *Cy/sm* to test whether they still carried sm^3. A recombinant stock, $Cy/L\ nw^B\ sm^3\ Pu$, was established, which was subsequently outcrossed with the M strain Samarkand (Lai and Mackay 1990). The final stock, LP/SAM, contained three or four P elements in the total genome, but only one on the right of the second chromosome (56E).

A genomic library was constructed as in Methods, except that instead of using *Sau*3AI for the partial digest, *Mbo*I was used (*Sau*3AI is an isoschizomer of *Mbo*I). The library was screened using pPiBWC (a gift from K. O'Hare), a plasmid which contained almost the entire P element, but no flanking sequences.

By virtue of P element homology, positive lambda clones (16 in total) were isolated, biotin labelled and used as probes for *in situ* hybridisations on Samarkand larvae (which do not contain any P elements). The only clone which hybridised to the 56E site was the subsequently-labelled λG1 clone. Restriction mapping of the λ clone indicated a probable full length insertion of the 2.9 kb P element. λG1 was subcloned and a number of different genomic libraries were constructed. Genomic DNA adjacent to the 3' and 5' end of the P element insertion site was used to probe the libraries listed below:

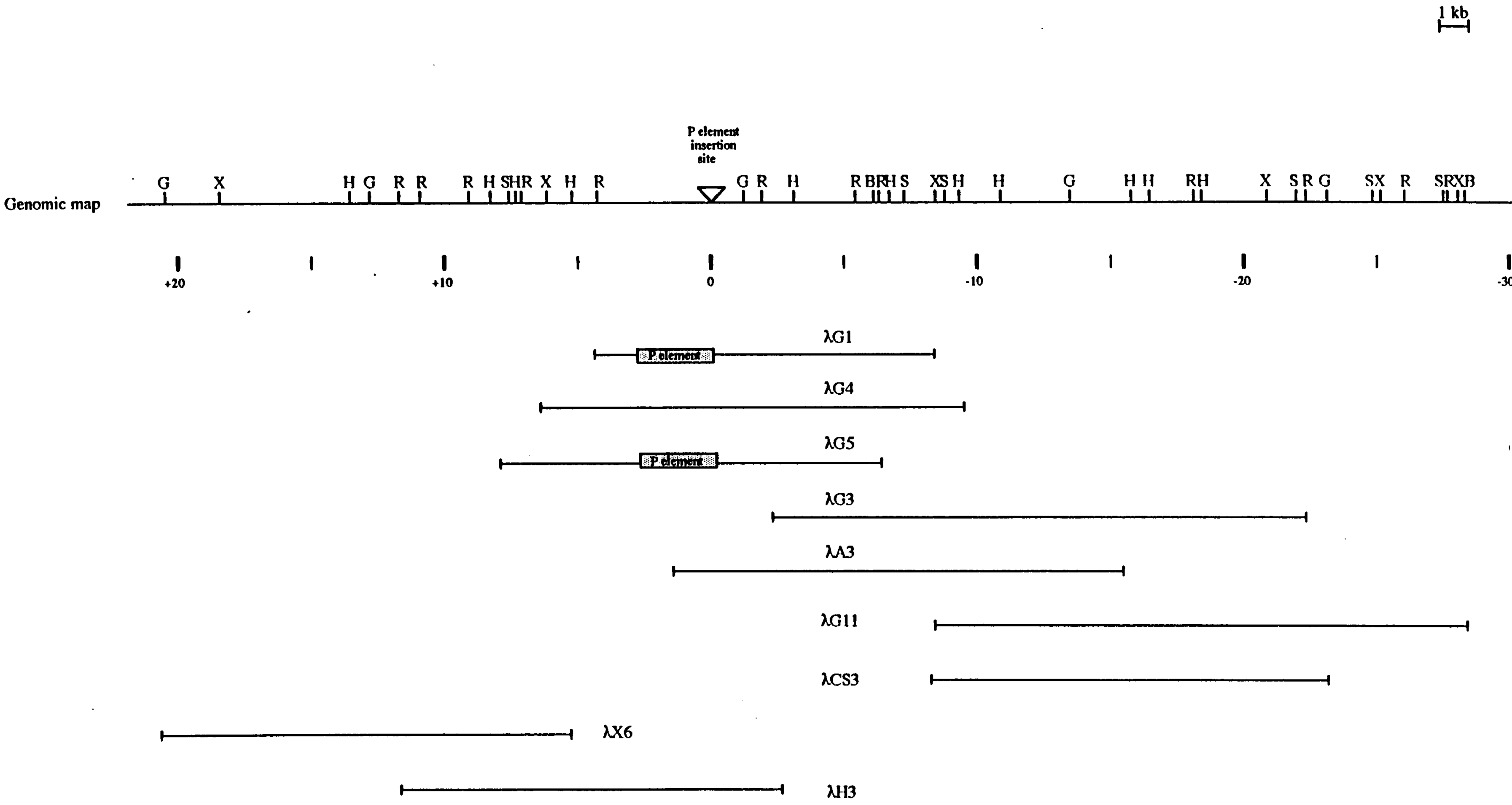
- LP/SAM (λG, original library as described above)
- SAM (λA, Samarkand, M strain, wild-type)
- Harwich (λH, P strain, wild-type)
- *Cy/sm px* (λX, heterozygous strain from stock centre)
- Canton-S (λCS, M strain, wild-type)

Figure 1 depicts the restriction maps of several of the overlapping λ clones isolated by A. Shrimpton showing the *Eco*RI (R), *Bam*HI (B), *Xho*I (X), *Bgl*II (G), *Hind*III (H), and *Sal*I (S) restriction sites only. A genomic region spanning 50 kb (illustrated as continuous horizontal line) was covered by the λ clones, with 20 kb downstream of the 3' end of the P element insertion site and a further 30 kb upstream of the 5' end of the P element. The coordinate 0 was fixed at the P element insertion site. The proximal-distal orientation of the restriction map with respect to the chromosome was not established.

Figure 1: Lambda maps

Figure 1 shows a horizontal line representing a genomic restriction map covering approximately 50 kb of the region surrounding the P element insertion (▽). The coordinates (in kb) are set zero at the point of P element insertion. Six different restriction sites are indicated as small vertical ticks on the map and labelled by single capital letters: *Eco*RI (E), *Bam*HI (B), *Hind*III (H), *Xho*I (X) and *Bgl*II (G).

Below the genomic map, nine overlapping lambda clones are depicted, covering the entire length of the genomic map. Genomic libraries from different *Drosophila* strains and lines were used to isolate those clones, as indicated in the text (λG LP/SAM; λA Samarkand; λH Harwich; λX *Cy/smpx*; λCS Canton-S).


1 kb
P element insertion site
Genomic map
G X H G R R R H SHR X H R G R H R BRH S XS H H G H H RH X S R G SX R SRXB
+20
+10
0
-10
-20
-30
λG1
P element
λG4
λG5
P element
λG3
λA3
λG11
λCS3
λX6
λH3

Subcloning of λ clones

A number of *Eco*RI subclones, ligated into the plasmids pUC8 and pGEM2, were prepared by A. Shrimpton. These subclones were predominantly derived from the following λ clones: λG1, λG3, λG5, λA3 and λX6 (see listing above). Many of these subclones were located upstream of the site of P element insertion, between position 0 and -10 on figure 1. Initially, the subclones served as probes for obtaining further lambda clones and for mapping purposes. The most frequently used subclones are depicted in figure 8 (p4s, p4r, p4g and p4d). All of these were derived from the wild-type Samarkand chromosome of λG4. In particular p4r, p4g and p4d were subsequently utilised as probes to examine the difference between the genomic DNA of the *smooth* alleles and the wild-type.

The start of the next section marks the work of the project I have carried out myself.

4.1.2 Analysis of the P element insertion site

The initial requirement for this project was the precise localisation of the P element insertion site in the recombinant lambda clone, λG5. In order to achieve this, the cloning of two terminal fragments, a 3.3 kb *Hind*III–*Hind*III fragment at the 5' end and a 4.8 kb *Eco*RI–*Sal*I fragment at the 3' end of the P element was carried out, as shown in figure 2. Subcloning of these clones followed and at the 5' end of the P element a 924 bp *Hind*III–*Hinc*II fragment, including 39 bp of P element at the *Hind*III site, was isolated and cloned into pTZ19R (*Hind*III/*Sma*I cut). A 685 bp *Hinc*II–*Sal*I fragment containing 497 bp of the 3' end of the P element was inserted into pTZ19R.

Figure 2: λG5 map

λG5 was derived from the LP chromosome containing the P element insertion at 56E IIR. The restriction map of λG5 shows the sites of eight restriction enzymes: *Eco*RI (R), *Bam*HI (B), *Hind*III (H), *Xho*I (X), *Bgl*II (G), SalI (S), *Pst*I (P) and *Hinc*II (C). The sites in brackets at either end of the λ clone are sites present only in the polylinker of the vector (EMBL4), not in the λ clone.

The P element in the centre of the map is depicted as a grey box, with an indication of the position of its 5' and 3' end. The initial large subclones from either end of the P element are shown: a 3.3 kb *Hind*III–*Hind*III fragment at the 5' end and a 4.8 kb *Eco*RI–*Sal*I fragment at the 3' end.

Both fragments were further subcloned: into a 924 bp *Hind*III–HincII at the 5' end and a 924 bp *Hinc*II–*Sal*I fragment at the 3' end of the P element. The *Hind*III–*Hinc*II fragment contains 39 bp of the 5' end of the P element and the *Hinc*II–*Sal*I fragment contains 497 bp from the 3' end.

Sequencing was carried out from either end of both subclones and the immediate sequence surrounding the 8 bp repeat on either site of the P element insertion is shown, with the 8 bp repeat sequence in underlined capital letters, the genomic sequence in capital letters and the P element sequence in small letters.

sm^{P}chromosome :λ G 5

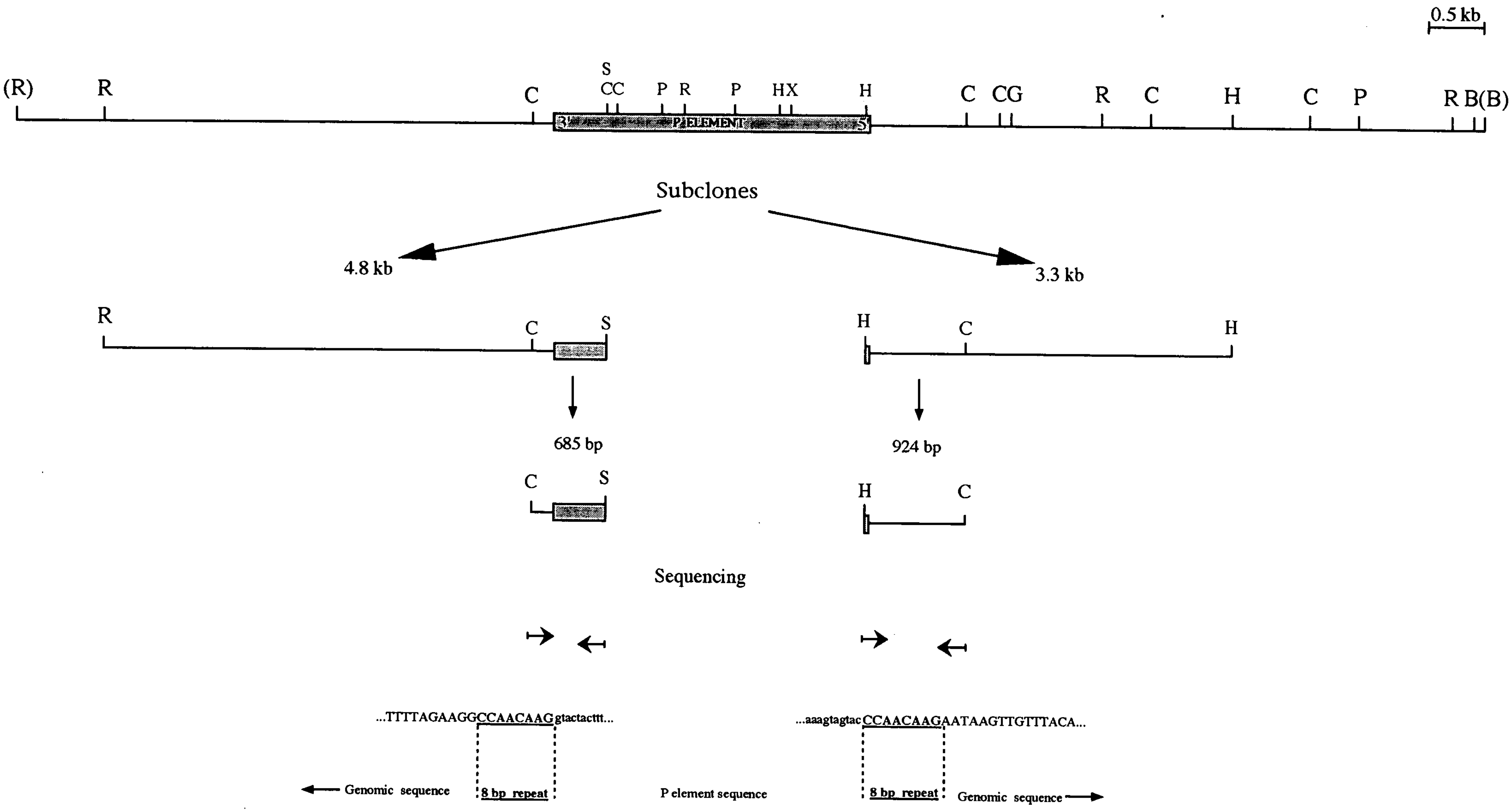


Figure 2 describes the cloning and subcloning of the restriction fragments from λG5 which include the 5' and 3' end of the P element together with the flanking genomic sequences.

Both subclones were sequenced for at least 250 bp from either end. The sequence confirmed a duplication of 8 bp of genomic DNA, **GAACAACC**, the so-called target sequence, on either site of the P element insert. This 8 bp duplication is thought to be a remainder of a staggered cut caused during the insertion of the P element. The 8 bp target sequence obtained matches the consensus sequence, **GGCCAGAC** (O'Hare and Rubin 1983), in 4 out of 8 base pairs.

The *Hinc*II fragment, into which the P element had inserted, was subsequently isolated from p4r (as described above) of λG4, Samarkand wild-type chromosome. The 1.1 kb *Hinc*II fragment was subcloned into pTZ19R (*Sma*I cut).

The whole *Hinc*II fragment of 1073 bp, fig.3, was sequenced by generating nested deletions (see Methods). It was suspected to include part of the coding sequences of the gene, since P elements have a tendency to insert at or near a transcription start site (Tsubota *et al.* 1985; Chia *et al.* 1986; Searles *et al.* 1986; Kelley *et al.*1987; Roiha *et al.* 1988).

At basepair 879 of the *Hinc*II fragment sequence, a single copy of the 8 bp target sequence was found. That site corresponds to the P element insertion site in λG5.

4.1.3 Isolation and characterisation of a cDNA clone

In order to identify potential transcripts of the *smooth* gene, genomic DNA clones (previously isolated by A. Shrimpton) situated upstream of the 5' end of the P element, p4g and p4d (shown in fig.8), were labelled and used to probe Northern blots. Neither of the clones revealed any hybridisation to the RNA filter (results not shown).

Figure 3: *Hinc*II fragment sequence

The complete *Hinc*II fragment sequence of 1073 bp is shown with the 8 bp repeat sequence, the site of P element insertion, underlined (from 879–886 bp).

Hinc II fragment sequence

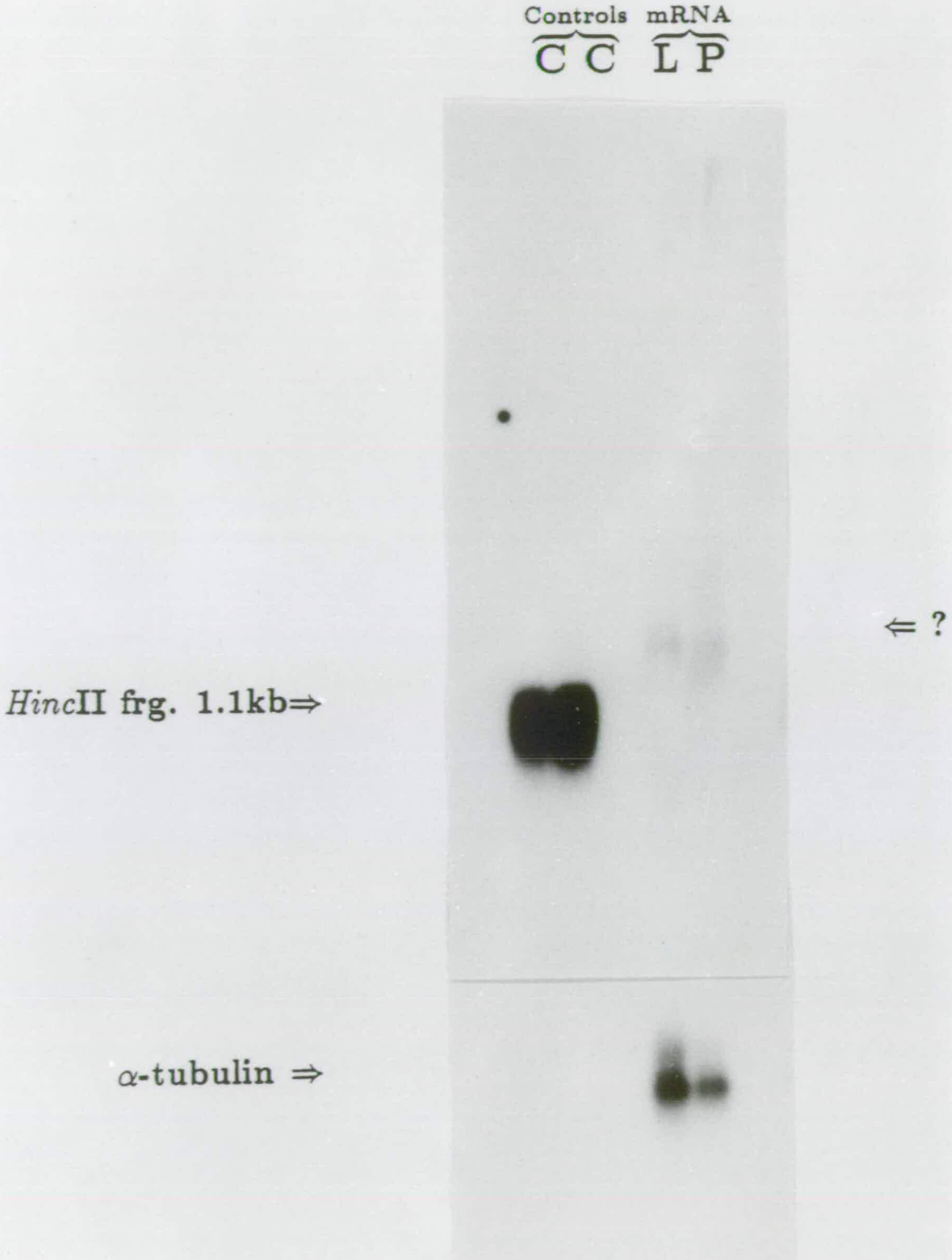
```
   1  caaccttattatatatatattattttcgtcgtcgcttattttcagctactgtttctcct   60
  61  ctattatattttatttgcttcgccattgtgtttaggatattttatttatttgagatagtg  120
 121  tttaaataagatttaataatatgttaaatttaaaacataaacatttttttgacagcaaat  180
 181  tttgccctcacaaagcaaagttttgaccacagtacgctgcctctgccgtcgctgccgctg  240
 241  agggagaaaaacgtatctgagttttgcttgggttgcgtttggttttcgcgtcagtttcga  300
 301  ggtttcgaggttcgagctaaaactttgtgcacggagcaaattaaaaataataaaataaga  360
 361  aataaaataataacaggcataagaaaagaaaagcgacgacaacaacaaaagctgatgaat  420
 421  acgtgcgttcgtgtgggtgccgtgcattctgcgtacatatttacaagcattcgtacgctt  480
 481  aaattaaatactttggcttaattaaatacttccgcggcgtcgttgttgtagttgttgata  540
 541  ttttgcctcttcacgcgttgtctacggttatcgattgtcgtcgtctcgctccctcgtacg  600
 601  catacgcatcgagttgttgtttgtttgttgtttgtgtctgtcctcgcgttttgtttgtgt  660
 661  actgacaaaataacaacaaaataaaacccgaaagactgcgaataagcataaataataaac  720
 721  aacgagacaagtgcaaaaaaagcgcagcgaaaaggccaggtaaaaaaaagaaaagctggc  780
 781  gctaacggtgaaagttgcgattgtgtgtgcaaatgaaaatccagtgtatgcgtgtgtgtg  840
 841  tgtgcaaggcagaaacagaaaccacatttgttgaataaGAACAACCggaagattttcct  900
 901  ttgccacccacgaaaattcagcacaaagggccaggcaaaaaggtagttctcgcgagagc  960
 961  gggagagagagtgagactgagagaattagagttgcaacactctctcctctctcttttcg  1020
1021  gacaccttgcaacccttaaagaaatgctctctctgcgccgaccgttagttggg  1073
```

Figure 4: Northern blot

This figure shows the photograph of a Northern blot with the first two lanes (C and C) being identical and only containing T7 transcripts of the 1073 bp *Hinc*II fragment which were used as controls. The other two tracks contain approximately 2 μg of larval (L) and pupal (P) poly(A)$^+$ RNA from the Samarkand *Drosophila* strain.

The blot was initially probed with a ^{32}P radiolabelled *Hinc*II fragment probe and subsequently reprobed with a 1.8 kb α-tubulin probe.

Opposite page 83


Controls mRNA
C C L P
⇐ ?
*Hinc*II frg. 1.1kb⇒
α-tubulin ⇒

Subsequently, the genomic 1.1 kb *Hinc*II fragment described above, which spans the insertion site of the P element at 56E, was used as a probe for Northern blots. The blots obtained were positive but ambiguous, with some Northerns showing a weak hybridisation band of bigger than 1.1 kb (fig. 4) whereas others show up two very weak transcripts of larger sizes. Despite the ambiguity, the HincII fragment was chosen to probe a cDNA library.

An amplified cDNA library prepared from 12–24 hr embryos (Brown and Kafatos 1988) was obtained. In this library, cDNA clones are directionally inserted into a specially designed plasmid vector. Approximately 120,000 colonies were screened, using the HincII fragment as a probe and 10 positives were obtained after the initial screening. Following repeated plating and purification steps, three positive cDNA clones were recovered and, since an amplified library had been screened, it was hardly surprising that all of them were of the same length when compared on an agarose gel. The cDNA clones were 2.6 kb long.

A previous attempt to isolate a positive cDNA clone had been unsuccessful. A different aliquot of the same amplified cDNA library was plated and 46,000 colonies were screened. The screening procedure was identical to the method used for the second attempt, as mentioned above and outlined in Methods (chapter 2).

As many as 32 positive colonies were observed after the first hybridisation, using the 1.1 kb *Hinc*II fragment as a probe. All of these colonies were isolated and replated. Eventually, positive clones were purified and characterisation was carried out by restriction mapping. The initial restriction digests already indicated a difference between the anticipated pattern according to the plasmid map of pNB40 (Brown and Kafatos 1988) and the pattern obtained.

To further analyse this discrepancy and possibly confirm a suspicion of contamination, sequencing reactions were carried out. Eventually, the "positive clone" was identified to be a contaminant, a subclone of the *Hinc*II fragment inserted into pTZ19R.

Since the contamination of the cDNA library stock (maintained as an ethanol suspension) could have taken place during the initial precipitation step of the original aliquot of the cDNA library, a fresh aliquot of the cDNA library was obtained to repeat the procedure. A number of precautions were taken in order to eliminate the chance of any further contamination.

To verify that the cDNA clones from the second screening attempt were derived from the *smooth* locus, one of the cDNAs, after having established most of its sequence, was biotin labelled and hybridised to polytene chromosomes of Samarkand larvae. The result of this *in situ* hybridisation is shown in photographs (fig.5) below, where clear hybridisation bands at the cytological position of 56E (IIR) are visible. This cDNA therefore derives from the site of insertion of the P element in the sm^3 mutation.

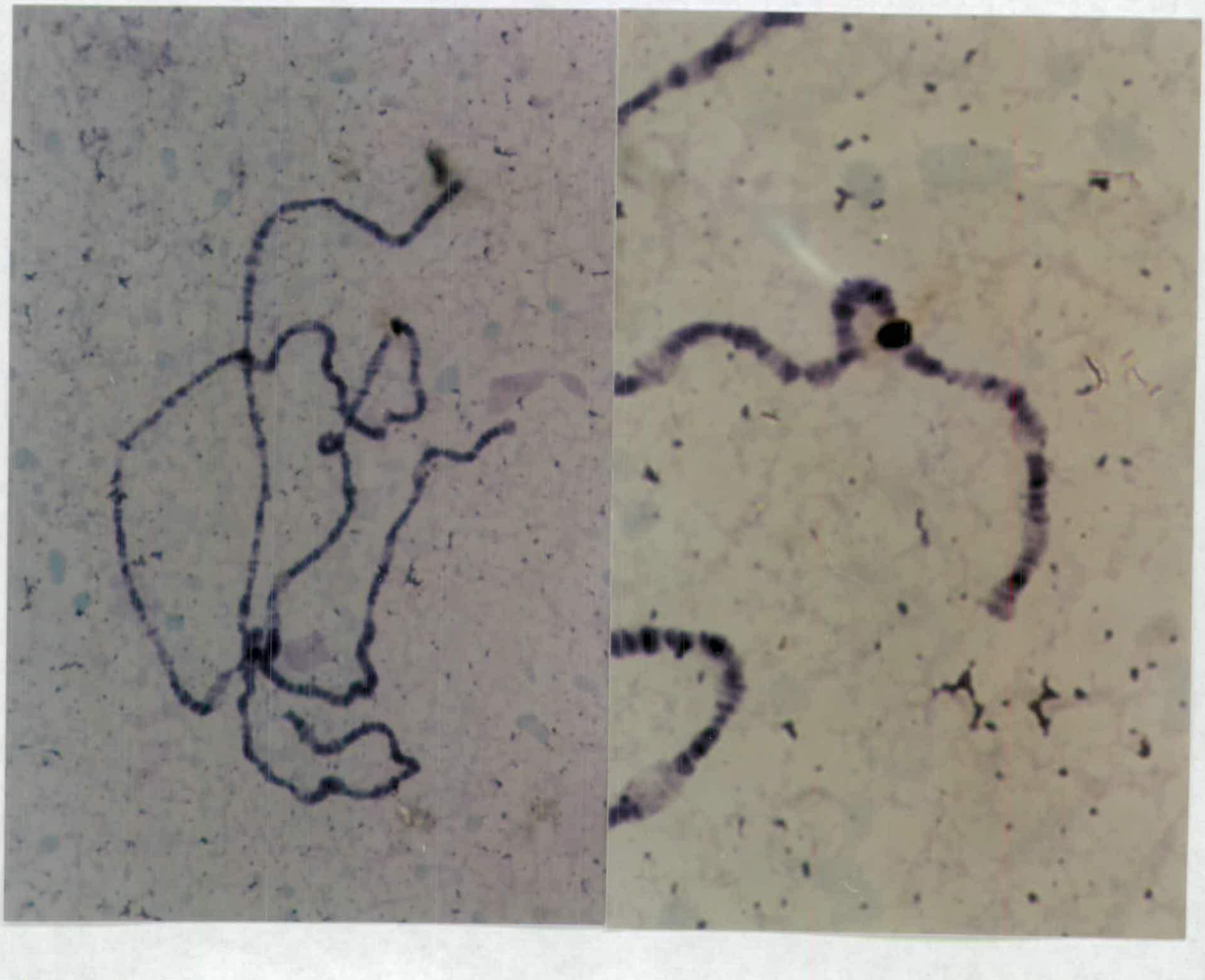


Figure 5: Photographs showing *in situ* hybridisation at two different magnifications (40× and 100×)

Restriction mapping of the cDNA clones with commonly used enzymes that recognise six basepair sequences (i.e. *Eco*RI, *Hind*III, *Pst*I, *Sal*I etc.) was carried out. Of these, the only restriction enzyme which cleaved the cDNA clone was *Pvu*II, which cuts twice, producing three fragments, of 2.9 kb, 1.2 kb and 0.9 kb in length. One *Pvu*II restriction site is present at the 5' end of the vector sequence at position 76 (fig.6).

The three restriction fragments of the cDNA were subsequently subcloned into pTZ19R and pTZ18R to facilitate the sequencing. The following restriction fragments were isolated, as indicated in figure 5:

1. subclone 1: the 5' *Hind*III–*Pvu*II fragment: 532 bp

2. subclone 2: the central small *Pvu*II fragment: 952 bp

3. subclone 3: the 3' large *Pvu*II fragment: 1271 bp

Fragments 1 and 3 contain 66 bp of 5' (from the *Hind*III site) and 76 bp of 3' (up to the *Pvu*II site) vector sequences, respectively.

4.1.4 cDNA sequencing

Sequencing was predominantly performed on either of the two strands, using the M13 primers (forward and reverse) in pTZ18R and pTZ19R on double–stranded DNA. Sequences which extended too far away from the M13 primers were accessed by designing 18–20 bp primers derived from a 3' end of a newly obtained part of the cDNA sequence. The range of sequence readable extended routinely upto 250–300 bp in one sequencing run. Everything was sequenced at least three times to ensure the accuracy of the sequence data. Also, when nucleotide sequences could not be resolved unambiguously, new primers close to the site in question were applied to the sequencing reaction.

Figure 6: cDNA plasmid

Figure 6 shows the map of the plasmid containing the cDNA. It was drawn using the PLASMIDMAP program of the UWGCG package on the VAX computer. The site and length of the vector pNB40 (2498 bp) is indicated by a long arrow inside the circle of the plasmid. Only the *Eco*RI sites (2), the *Hind*III site and the *Pvu*II site of the vector are shown. The cDNA (2613 bp) insert is emphasised by a thicker bar with a criss-cross pattern.

The *Pvu*II subcloning sites of the cDNA are shown with their base pair positions in brackets. The lines inside the circle of the plasmid labelled 1,2 and 3 covering the whole length of the cDNA and short sequences at either end of the pNB40 vector are the three subclones:

Subclone 1: *Hind*III–*Pvu*II fragment

Subclone 2: central small *Pvu*II–*Pvu*II fragment

Subclone 3: large *Pvu*II–*Pvu*II fragment.

Subclone 1 contains 66 bp of the 5' end of the pNB40, starting at the *Hind*III site and subclone 3 includes 76 bp of the 3' end of the vector, up to the *Pvu*II site.

Opposite page 87

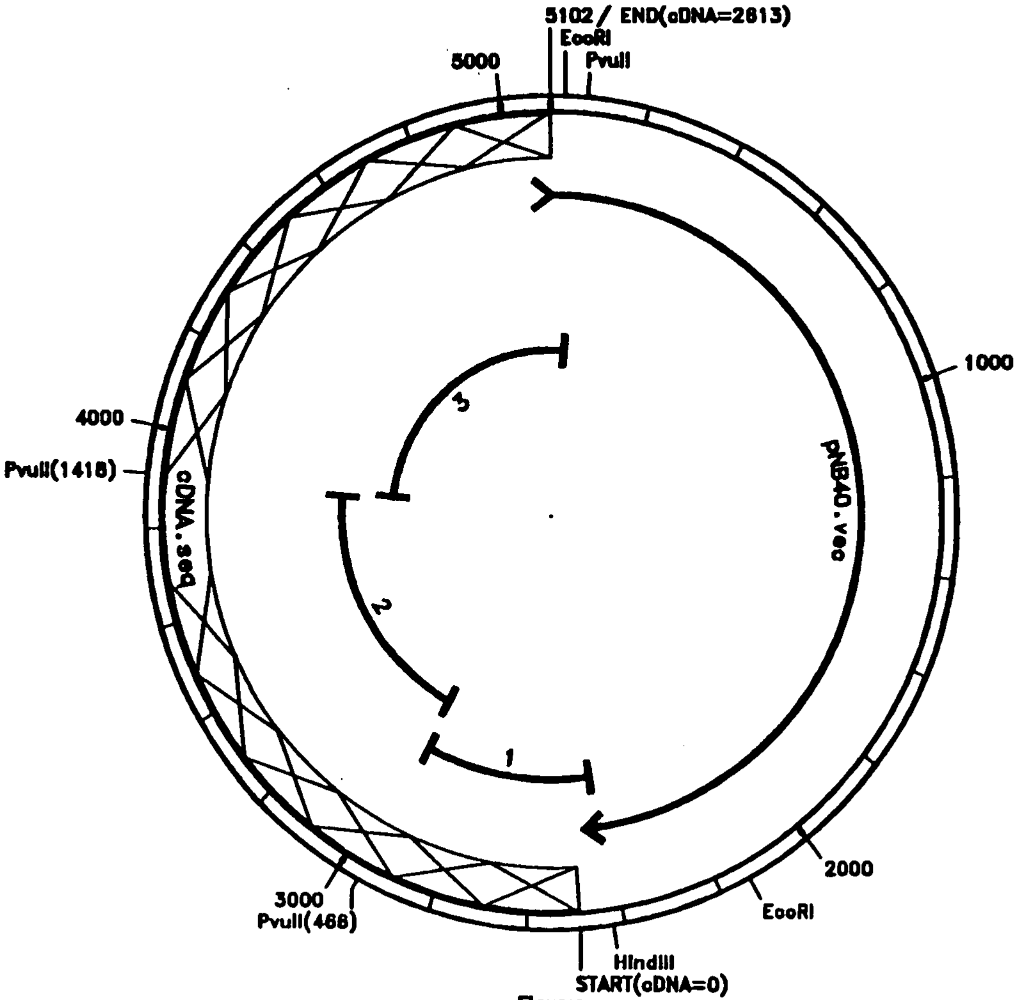


Figure:
cDNA clone (2813 bp) and vector (2489 bp)

All the parts of the sequence were assembled using the University of Wisconsin GCG programs (Devereux et al. 1984). The complete nucleotide sequence of 2613 bp, including an indication of the primer positions, is shown in the Appendix.

4.2 The organisation of the gene

4.2.1 Exon mapping

Exons were determined by comparison of cDNA and genomic sequences. Initially, the three above-mentioned cDNA subclones were used as probes, labelled by random priming, to hybridise to genomic Southerns of wild-type Samarkand DNA and furthermore to a number of the λ clones which were available for the region (see fig.1). From the results obtained on the autoradiographs, a complex picture emerged with a large number of bands hybridising to the two *Pvu*II fragments (2 and 3). The autoradiograph of a genomic Southern (see fig.7) shows the bands hybridising to the large *Pvu*II fragment (subclone 3).

Isolation of exon 1 and exon2

The first two exons were quickly identified, since they were both found to be contained in the p4r fragment of λG4 (fig.8). The first exon of 466 bp hybridised fully to the 532 bp 5' subclone of the cDNA (subclone 1 which also includes the 66 bp of vector sequence at 5' end). Subclone 1 of the cDNA was found within the *Hinc*II fragment, extending from nucleotide 293 to 759 of the total sequence of 1073 bp (fig.9).

The comparison of the exon1 cDNA sequence and the *Hinc*II sequence disclosed three differences. It has to be mentioned that different *Drosophila* strains were

Figure 7: Genomic Southern

Figure 7 shows a photograph of a Southern blot. Approximately 1μg of genomic Samarkand DNA was digested with the restriction enzyme(s) as indicated above each of the ten tracks (*Eco*RI (R), *Sal*I (S), *Bgl*II (G), *Bam*HI (B), *Xho*I (X), *Pst*I (P)). The samples were subsequently loaded on a 0.6% agarose gel. As size markers, λ*Hind*III and λ*Pst*I digests were loaded on either site of the genomic digests (only the sizes shown here).

The Southern blot was first hybridised to a radiolabelled ^{32}P subclone 3 (large *Pvu*II–*Pvu*II fragment) probe and subsequently reprobed with a radiolabelled λ probe.

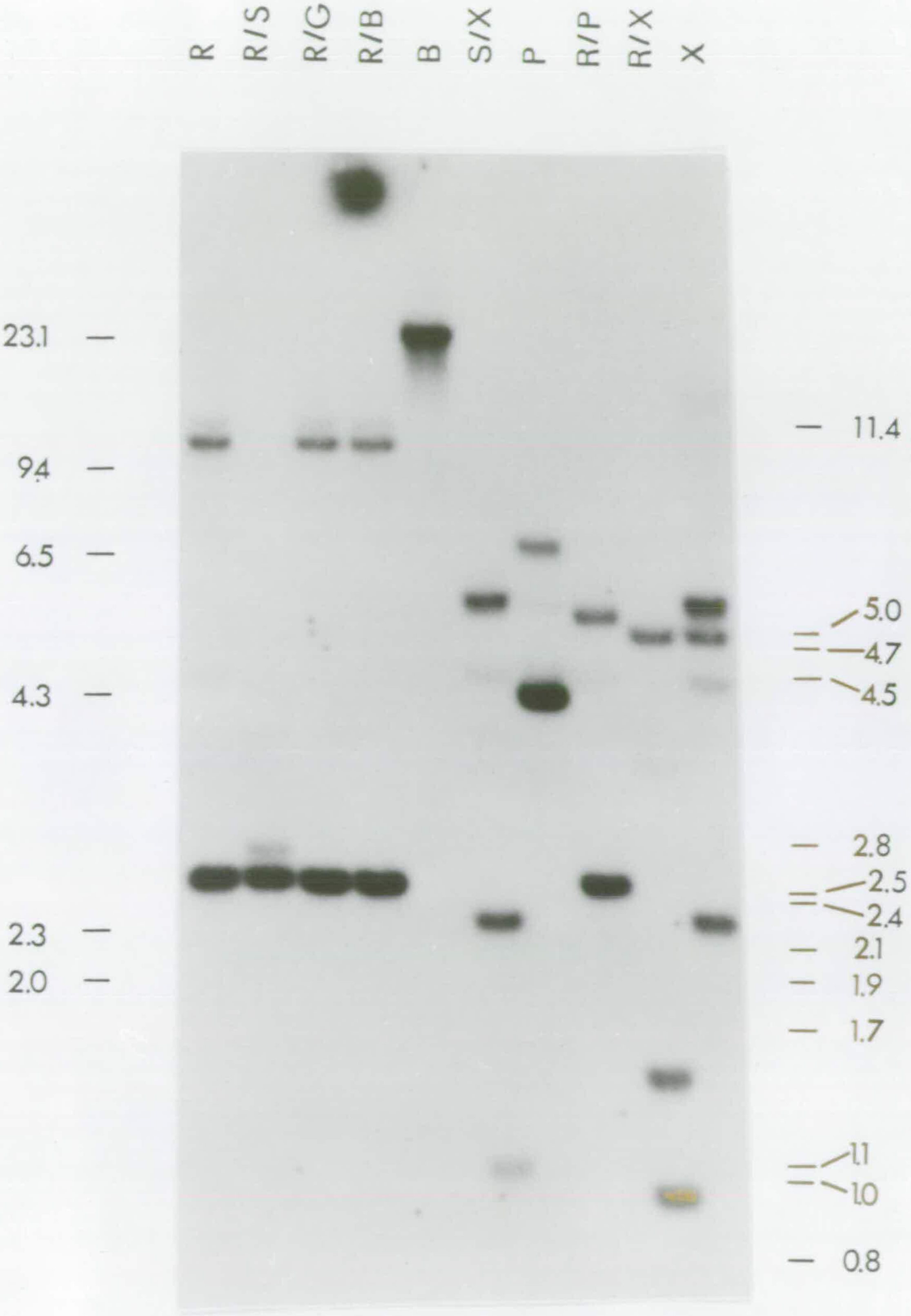

R
R/S
R/G
R/B
B
S/X
P
R/P
R/X
X
23.1
9.4
6.5
4.3
2.3
2.0
11.4
5.0
4.7
4.5
2.8
2.5
2.4
2.1
1.9
1.7
1.1
1.0
0.8

Figure 8: λG4 map

λG4 was derived from the *Drosophila* Samarkand genomic library. The long horizontal line shows the restriction map of λG4 with the following restriction enzymes: *Eco*RI (E), *Bam*HI (B), *Xho*I (X), *Hind*III (H), *Bgl*II (G), *Sal*I, *Hinc*II (C) and *Pst*I (P) sites. The *Eco*RI sites indicated in brackets at either end are part of the EMBL4 polylinker.

Four *Eco*RI subclones, here depicted as "probes", are shown (p4s, p4r, p4g and p4d) above the λG4 restriction map.

The 1.8 kb *Pst*I–*Pst*I fragment and the 1.1 kb *Hinc*II– *Hinc*II fragment shown below the restriction map as "subclones" were both derived from p4r (see also fig.9). The smaller of the two clones, the *Hinc*II fragment, was completely sequenced (see fig.3), and the *Pst*I–*Pst*I fragment was partially sequenced starting from either end.

The *Hinc*II fragment contains exon1 and the *Pst*I–*Pst*I fragment contains exon 2. All exon/intron junction sequences, including the splice sites, are shown.

wild-type SAM chromosome : λG4

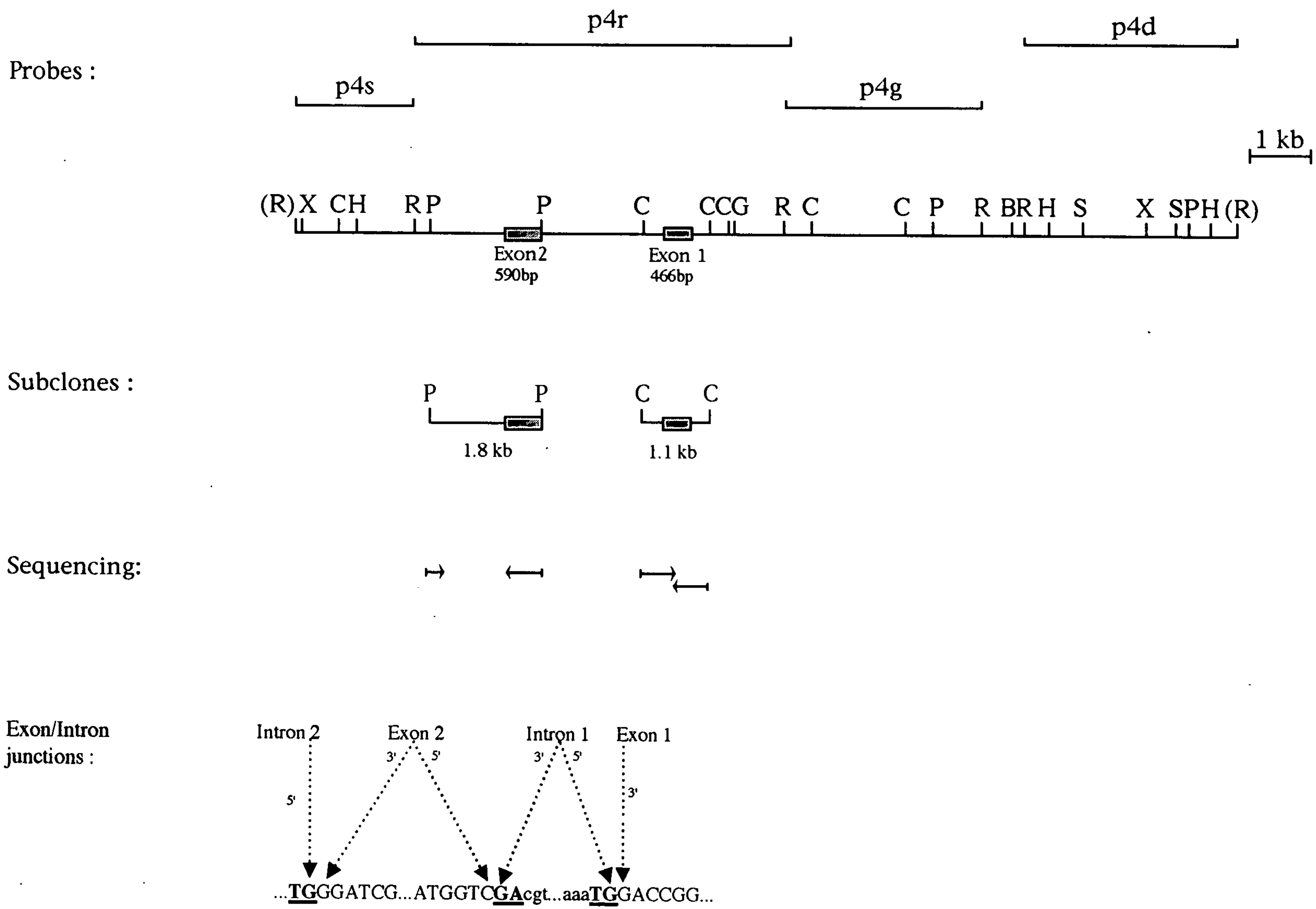
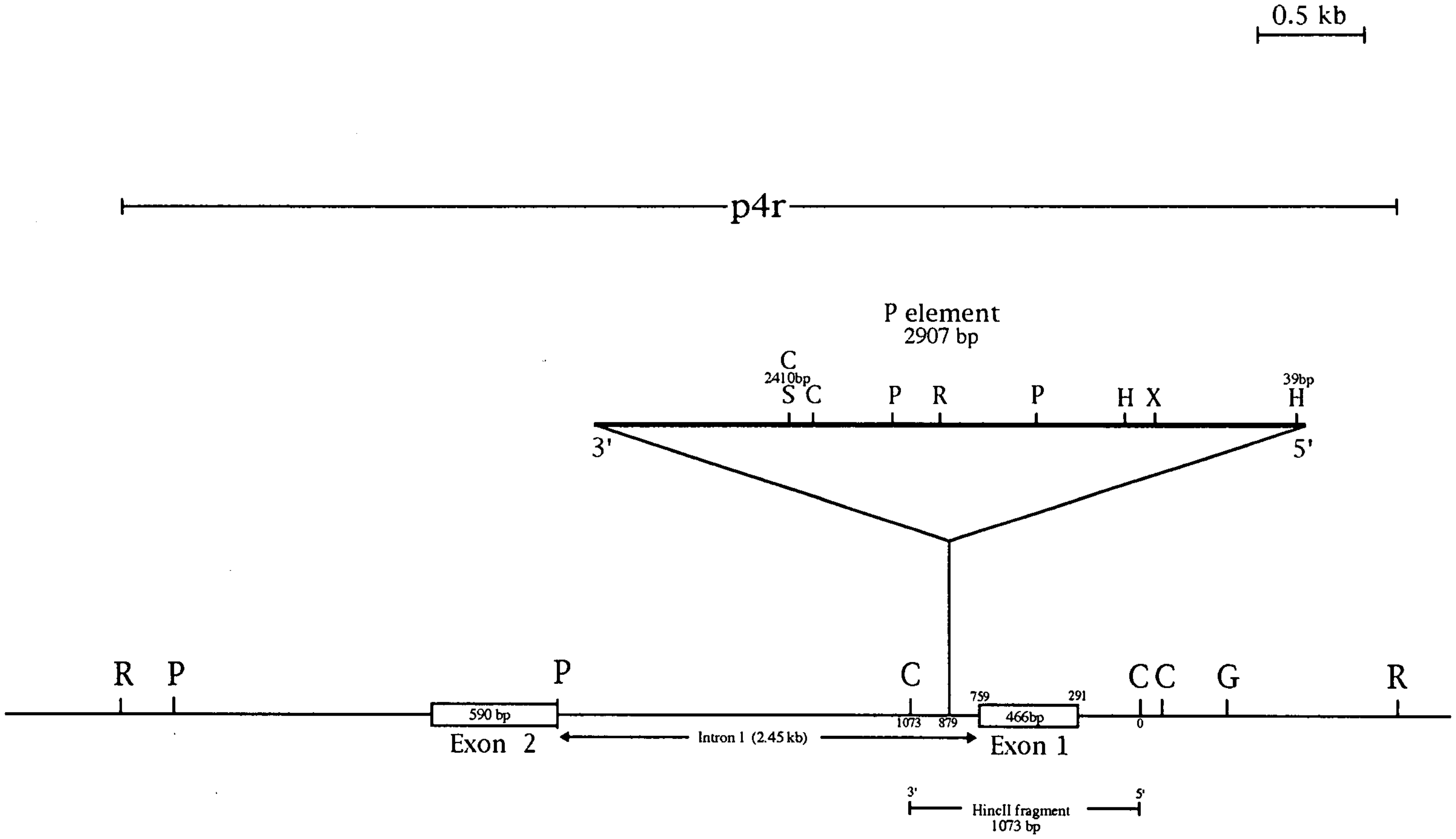


Figure 9: p4r including exons 1 and 2

This figure contains a more detailed description of figure 8. It also indicates the site of P element insertion (which is normally absent from the wild-type Samarkand chromosome) relative to the position of exon 1 and exon 2.

The p4r λG4 subclone is shown with the *Eco*RI (E), *Pst*I (P), *Hinc*II (C), *Bgl*II (G) restriction sites. The *Xho*I (X) and the *Sal*I (S) restriction sites are only shown for the P element. The size of the first exon, first intron and second exon, as well as the *Hinc*II fragment and the position of exon 1 in the *Hinc*II fragment, is shown.

Opposite page 91


0.5 kb
p4r
P element
2907 bp
C
2410bp
S C
P
R
P
H X
39bp
H
3'
5'
R P
P
C
C C
G
R
590 bp
Exon 2
Intron 1 (2.45 kb)
1073
879
759
466bp
291
0
Exon 1
3'
5'
HincII fragment
1073 bp

used for the cDNA library (an isogenic second chromosome strain carrying the markers *dp cn bw*) and the analysis (wild-type Samarkand strain).

First of all, the first nucleotide (G) of the cDNA sequence is not identical to what would be the first nucleotide (C) at the equivalent site of the *Hinc*II fragment. There are at least three possible explanations for this: The occurrence of a sequencing error can be ruled out immediately, since the sequencing was repeated several times. The presence of a polymorphism could be a second reason, but again the chance of the occurrence of a polymorphism at the first position of the cDNA must be very low. The third reason, and the most likely possibility could be that the G is an artifact. As explained under section 4.3.1 ("The putative transcript"), the 5' end G residue might have been incorporated during the reverse transcription reaction, when the cDNA library was made.

A second dissimilarity between the cDNA exon1 sequence and the HincII fragment sequence is a single nucleotide change. At position 602 of the *Hinc*II fragment is an A residue and, at the corresponding position 311 of the exon 1, a C residue is present. This single base pair change is probably a polymorphism.

Thirdly, a polymorphism showing an insertion of two additional A nucleotides was found in a run of seven A residues between position 735 and 741 of the *Hinc*II fragment. The equivalent site of the cDNA (444-448) only shows a run at 5 As. Sequencing error must be ruled out.

As it will be shown later, none of these three differences between the genomic and the cDNA sequences will affect the size of the transcript.

The intron separating the first exon from the second was 2.45 kb in length. The second exon, isolated by hybridising subclone 2 (the central *Pvu*II fragment of the cDNA) to λG4, consisted of 590 bp of nucleotide sequence and was contained in a 1.8 kb *Pst*I-*Pst*I genomic fragment. The 5' splice site coincides exactly with the upstream *Pst*I restriction site. Exon/intron junctions of both exons are shown in fig.8 and 9.

Having established the first exon positions, the orientation of the gene with respect to the genomic map became apparent (since the orientation of the cDNA in the vector was determined by its design). The gene extended in the same direction as the P element would be transcribed (5'→3'). This implied that all the λ clones 5' to the first exon were not required in the exon mapping. It gradually emerged, by probing the 3' λ clones, that none of the other exons was contained within the 50 kb genomic region covered by the λ clones isolated earlier by A. Shrimpton.

Due to the way the initial λ map (fig.1) was drawn, the 5'→3' direction in all the following diagrams illustrating the positions of the exons present, proceeds from right to left. To characterise further the structure of the gene, a different approach for isolating the exon containing fragments had to be applied: screening of genomic libraries.

4.2.2 Exons 3-10

A genomic library of Canton-S flies in EMBL4 (as described in Materials) was prepared and screened with the two *Pvu*II fragment subclones (2+3). A number of λ clones were isolated with subclone 3. Screening with subclone 2 was more difficult, since half of the subclone had already been localised within the 1.8 kb *Pst*I–*Pst*I fragment of p4r. Repeatedly, λ clones were isolated, which only covered the λG4 area. Therefore, a new library was obtained, an Oregon-R library in GEM–11 (gift of K. Kaiser), and by screening it, all the remaining exon positions were located.

Exon 3

Screening the Oregon-R library with the complete subclone 2, (the smaller, central *Pvu*II fragment of the cDNA), a large number of positives was isolated (at least nine). After further purification, filters were hybridised to the *Pst*I-*Pst*I fragment

Figure 10: λSP14 and λSP17 map

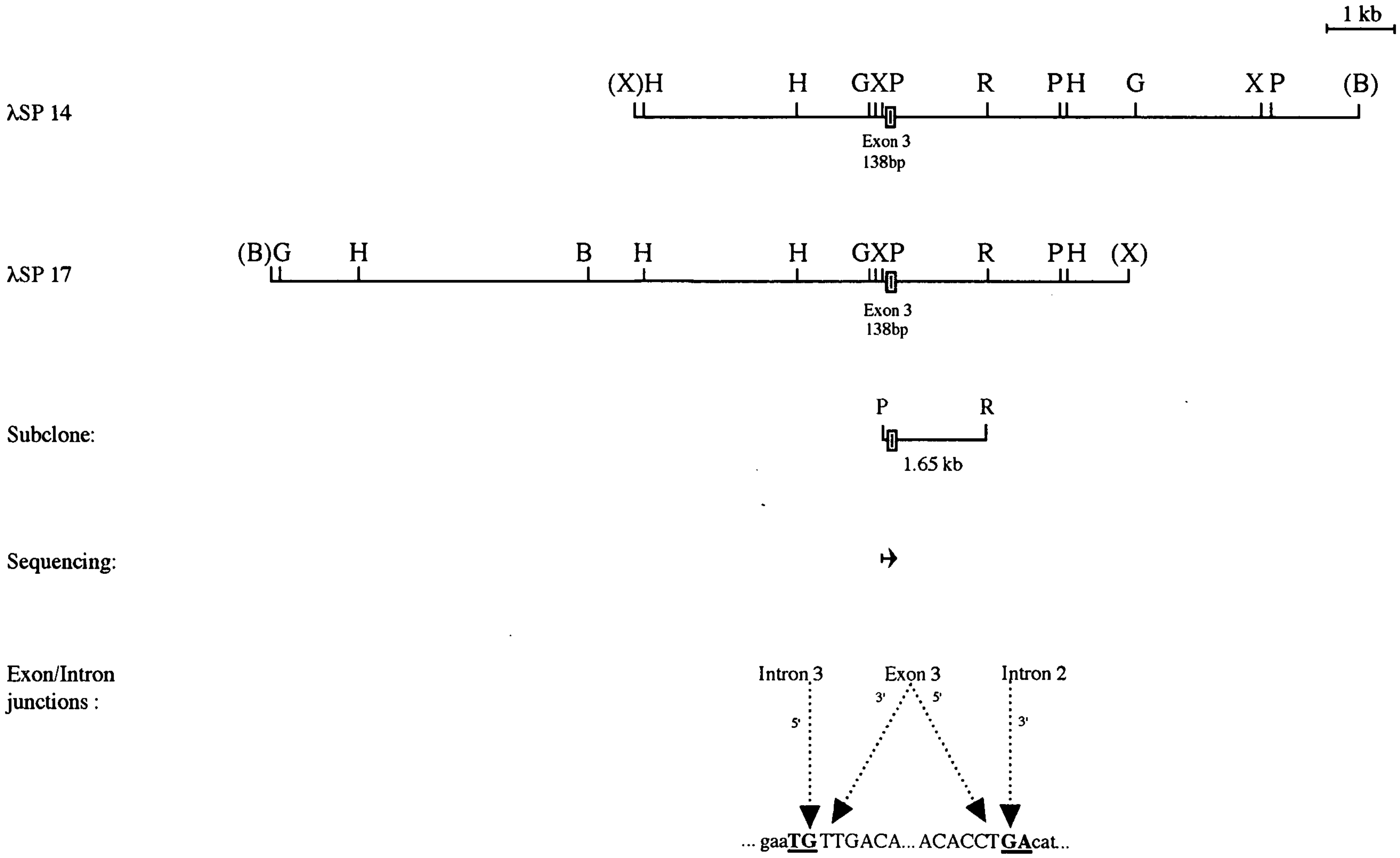
This figure illustrates the restriction maps of two overlapping λ clones, λ SP14 and λSP17, with the following restriction sites: *Eco*RI (R), *Hind*III (H), *PstI* (P), *Bgl*II (G) and *Bam*HI (B), where the *Bam*HI and the *Xho*I sites shown in brackets are part of the GEM-11 polylinker.

These λ clones had been isolated from the Oregon-R library probed with subclone 2 of the cDNA. Both λ clones contain exon 3 which was isolated by subcloning a 1.65 kb *Pst*I–*Eco*RI fragment. Sequencing was carried out from the *Pst*I side only as indicated by the arrows.

The sequences of the intron 2/exon 3 and exon 3/intron 3 junctions are shown.

Opposite page 94

wild-type chromosome: λ SP 14 and λSP 17



to eliminate all those positives which hybridised to the first 590 bp of the subclone. Eventually, two λ clones, SP14 and SP17, which did not hybridise to the *Pst*I-*Pst*I fragment were singled out.

Minilysate DNA of λSP14 and λSP17 was prepared and digested. Hybridisation of subclone 2 to Southern blots localised the homology to a *Pst*I–*Eco*RI fragment of 1.65 kb, as seen in figure 10. Subcloning into pTZ18R/19R was performed and initial sequencing from either end of the insert revealed 138 bp of sequence located at the cDNA 3' *Pst*I end of the fragment. Intron/exon junctions are depicted in fig.10.

The λ clones were restriction mapped and an overlap with λX6, the most 3' λ clone of the original set (fig. 1), was established by probing λSP14 and λSP17 with the 2.5 kb *BamHI*-*Xho*I (BX) probe derived from the 3' end of λX6 (fig.11).

Exon 4 and exon 5

Only 728 bp of the 952 bp of the small *Pvu*II cDNA fragment had been mapped so far, corresponding to 1194 bp of the complete cDNA. A 20 bp primer (646) was designed which started at position 737 bp of the small *PvuII* fragment (at position 1203→1222 bp of the complete cDNA). By the polymerase chain reaction (PCR), a 250 bp fragment was amplified from subclone 2 with the M13 (annealing to pTZ18R and 646 as primers, see below). This fragment was random-primer-labelled and hybridised to the Oregon-R library filter.

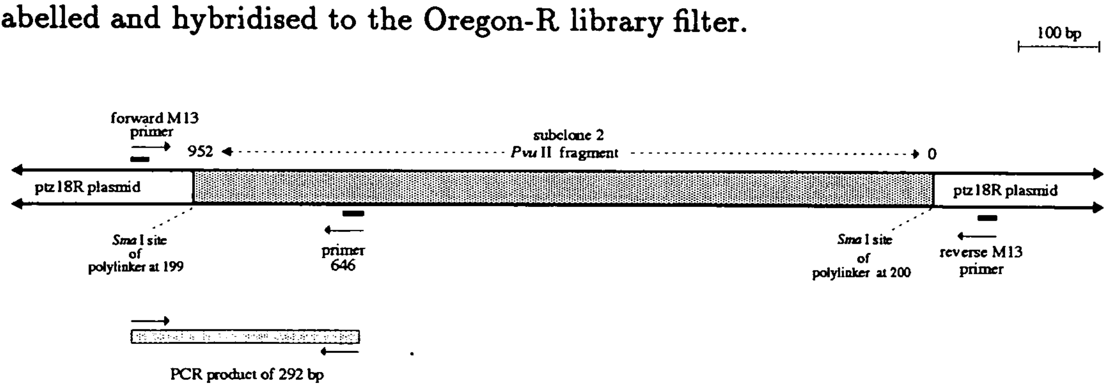


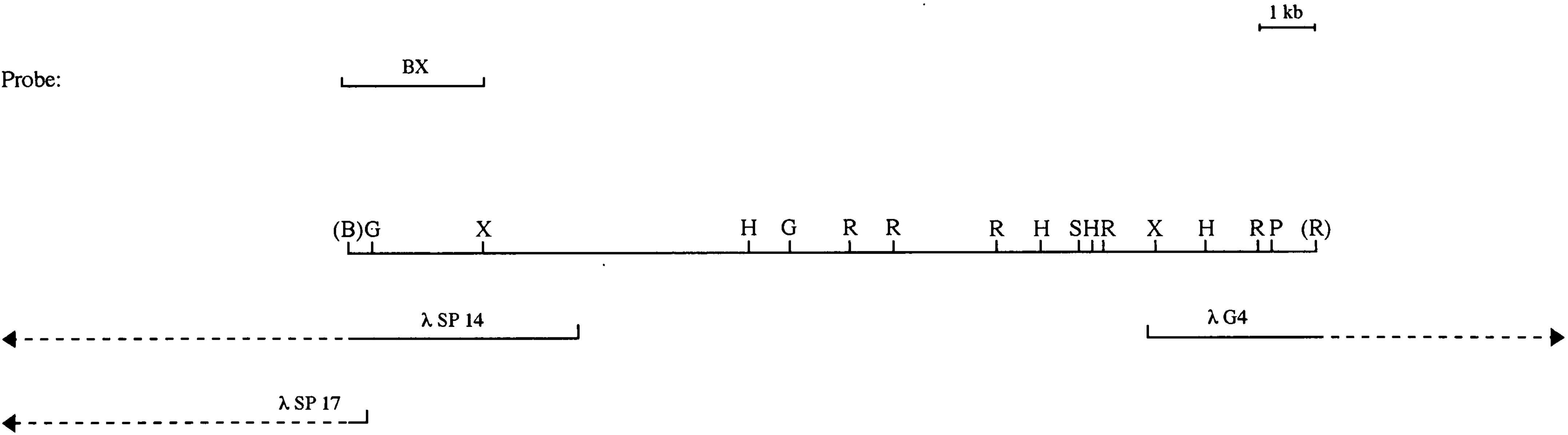
Figure 12: pTZ18R including subclone 2 insert

Figure 11: λX6 map

Figure 11 shows the map of λX6 with the following restriction sites: *Eco*RI (R), *Hind*III (H), *Sal*I (S), *Xho*I (X), *Bgl*II (G) and *Pst*I (P). The *Eco*RI and the *Bam*HI sites indicated in brackets are part of the polylinker of the EMBL4 vector.

λX6 does not contain any exon sequences, but overlaps with λG4 at its 5' end and λSP14 and λSP17 at its 3' end. A *Bam*HI–*Xho*I probe from the 3' end (as indicated) was used to establish the overlap with λSP14 and λSP17.

wild-type chromosone: λ X6



One very strong positive clone (λA22) was isolated, DNA was prepared and probed with the 250 bp fragment. The smallest single band to which the probe hybridised was a 0.75 kb *Xho*I-*Eco*RI fragment. This fragment was subcloned by first cloning the 2.7 kb *Eco*RI-*Eco*RI fragment into pTZ18R, as shown in figure 13, and subsequently subcloning the *Xho*I-*Eco*RI fragment into pBluescript II (KS-) (Stratagene).

The pBluescript II (KS-) plasmid contains a number of primer annealing sites located either inside or flanking the multiple cloning site. Sequencing of the *Xho*I-*Eco*RI fragment was therefore carried out, using the T3 and the SK primers which were located at either end of the insert. The 646 primer was also available for sequencing, since its sequence was present in the *Eco*RI-*Xho*I fragment.

A sequence of 317 bp obtained from the *Eco*RI 5' end did not overlap with the cDNA sequence, but sequencing from the 3' *Xho*I end and with the 646 primer revealed two exons separated by a 78 bp intron. Since the 646 primer derived from the cDNA sequence, started amplifying 8 bp from within the 5' end of exon 4, the intron 3/exon 4 junction was not confirmed.

No overlap between λA22 and the λ clones containing adjacent exons, i.e. λSP17 and λRX19 (see fig.14) was established.

Exon 6 and exon 7

Exon 6 and 7 were contained in a *Sal*I-*Xho*I fragment of λC (fig.14). The λC clone was isolated from the Canton-S library by probing it with the large PvuII fragment of the cDNA. The 1.1 kb *Sal*I-*Hind*III fragment was subcloned into pTZ18R and pTZ19R and was sequenced using the M13 primer starting from the *Sal*I 5' end of exon 6 (182 bp).

Figure 13: λA22 map

This figure illustrates the restriction map of λA22, containing exons 4 and 5. The positions of the following restriction sites were established: *Eco*RI (R), *Bam*HI (B), *Sal*I (S), *Xho*I (X) and *Bgl*II (G). The *Bam*HI site and the *Sal*I site surrounded by brackets are only contained in the polylinker of GEM-11.

A 2.7 kb *Eco*RI–*Eco*RI fragment was first cloned and a subclone, the 0.75 kb *Xho*I–*Eco*RI fragment was prepared from it. Sequencing was carried out, as indicated by the arrows, from either side of the subclone and the intron and exon junction sequences are shown.

Opposite page 98

wild-type chromosome: λA 22

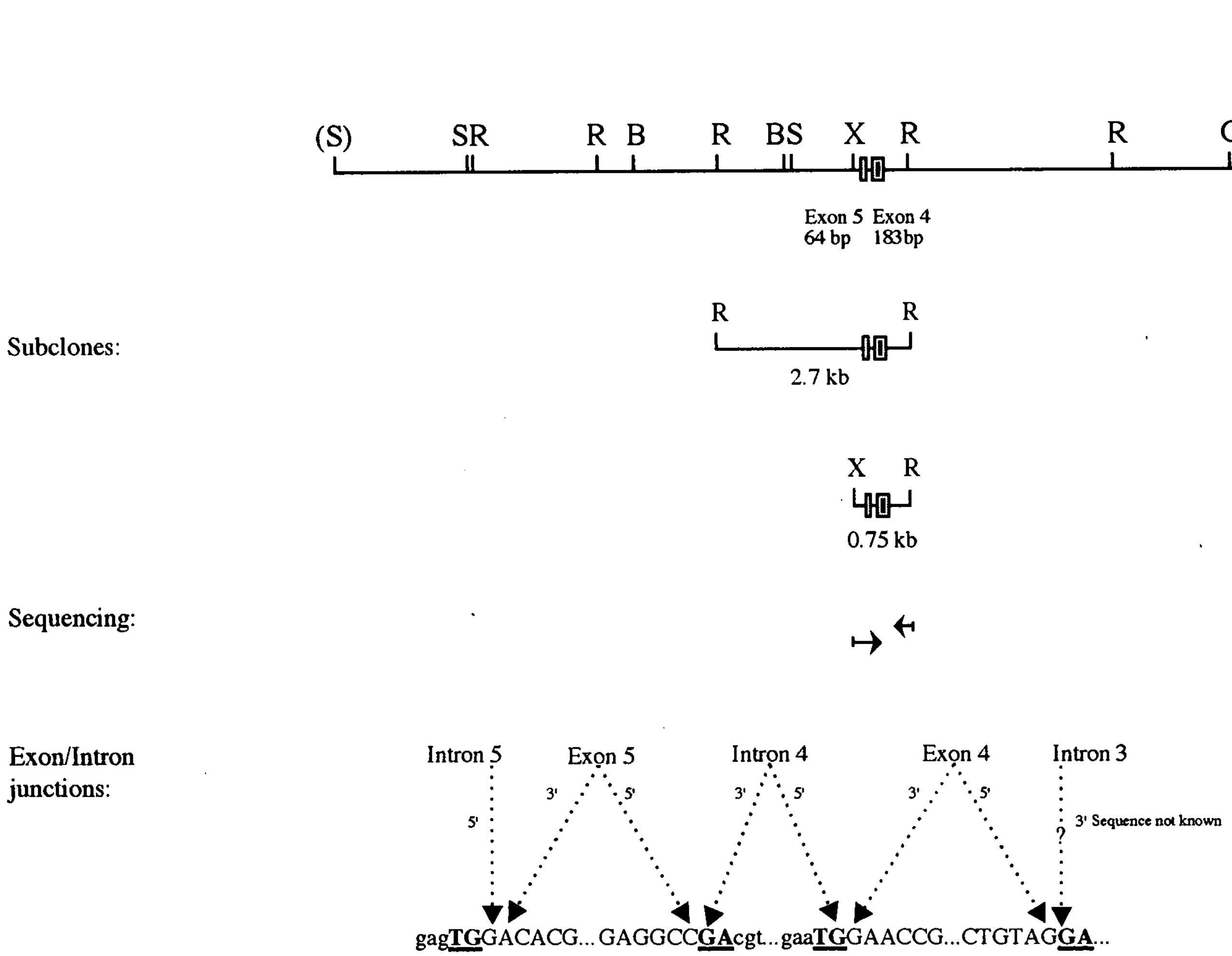


Figure 14: λRX19/λC map

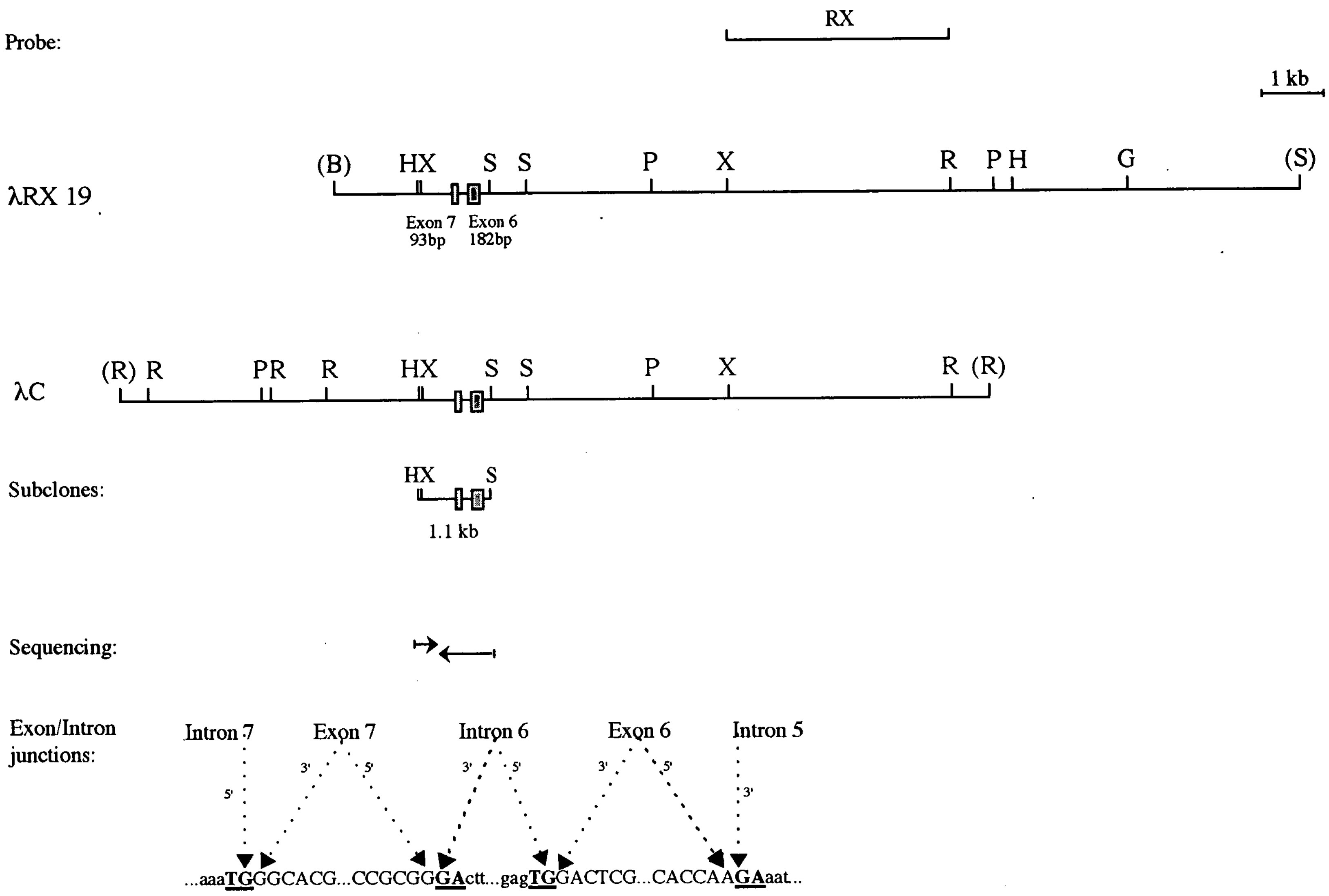
Figure 14 shows the restriction maps of two lambda clones, λRX19 and λC. The restriction enzyme sites established are *Eco*RI (R), *Hind*III (H), *Sal*I (S), *Bgl*II (G), *Xho*I (X) and *Pst*I (P). The restriction sites at the end of the lambda clones shown in brackets are part of the lambda vector polylinkers.

λC was derived from a Canton-S library and λRX19 from the Oregon-R library. The λC fragment was isolated before the λRX19 clone and, in an effort to carry out a chromosome walk to establish overlaps between the lambda clones, the 5' RX fragment (as indicated above the λRX19 map clone) of λC was used to probe the Oregon-R library and isolate λRX19.

A 1.1 kb *Hind*III–*Sal*I fragment was subcloned from λC and sequenced from either end. Two exons, exon 6 and 7 were found to be contained in the subclone. Exon and intron junctions were established as shown in the figure.

Opposite page 99

wild-type chromosome:λRX 19 / λ C



When primer 141 (20 bp long) was used for sequencing on the *SalI*-*Hind*III fragment, annealing 20 bp before the start of exon 6, an intron of 155 bp was detected followed by exon 7. The exon of 93 bp was sequenced, using primer 710 from 1647 to 1665 bp of the cDNA. All intron and exon junctions were established.

λRX19 also contains both exons and was isolated in an effort to find an overlap to a 5' λ clone. The most 5' 3.65 kb *Xho*I-*Eco*RI (RX) (fig.10) fragment of λC was used as the probe in the screening. But λRX19 did not provide the desired overlap with λA22 (fig.13).

Exon 8, 9 and 10

Although the complete exon 8 and parts of exon 9 were contained at the 3' end of λC, the Oregon-R library was screened again, using the 0.9 kb PCR product of pTZ18R plus cDNA insert with the reverse M13 primer and primer 824 (19mer, from position 1725 to 1743 of cDNA), which had originally been designed for the cDNA sequencing. Two identical clones, λA and λB, were independently isolated and minilysate DNA was prepared. Restriction analysis and Southern blotting indicated that the PCR product was hybridising to a 2.6 kb *Eco*RI-*Eco*RI fragment, which was subsequently subcloned into pTZ18R. When this fragment was further digested with the restriction enzyme *Xho*I, producing a 0.95 kb and 1.65 kb fragment, hybridisation with the above mentioned probe was found to occur to both fragments (fig.15).

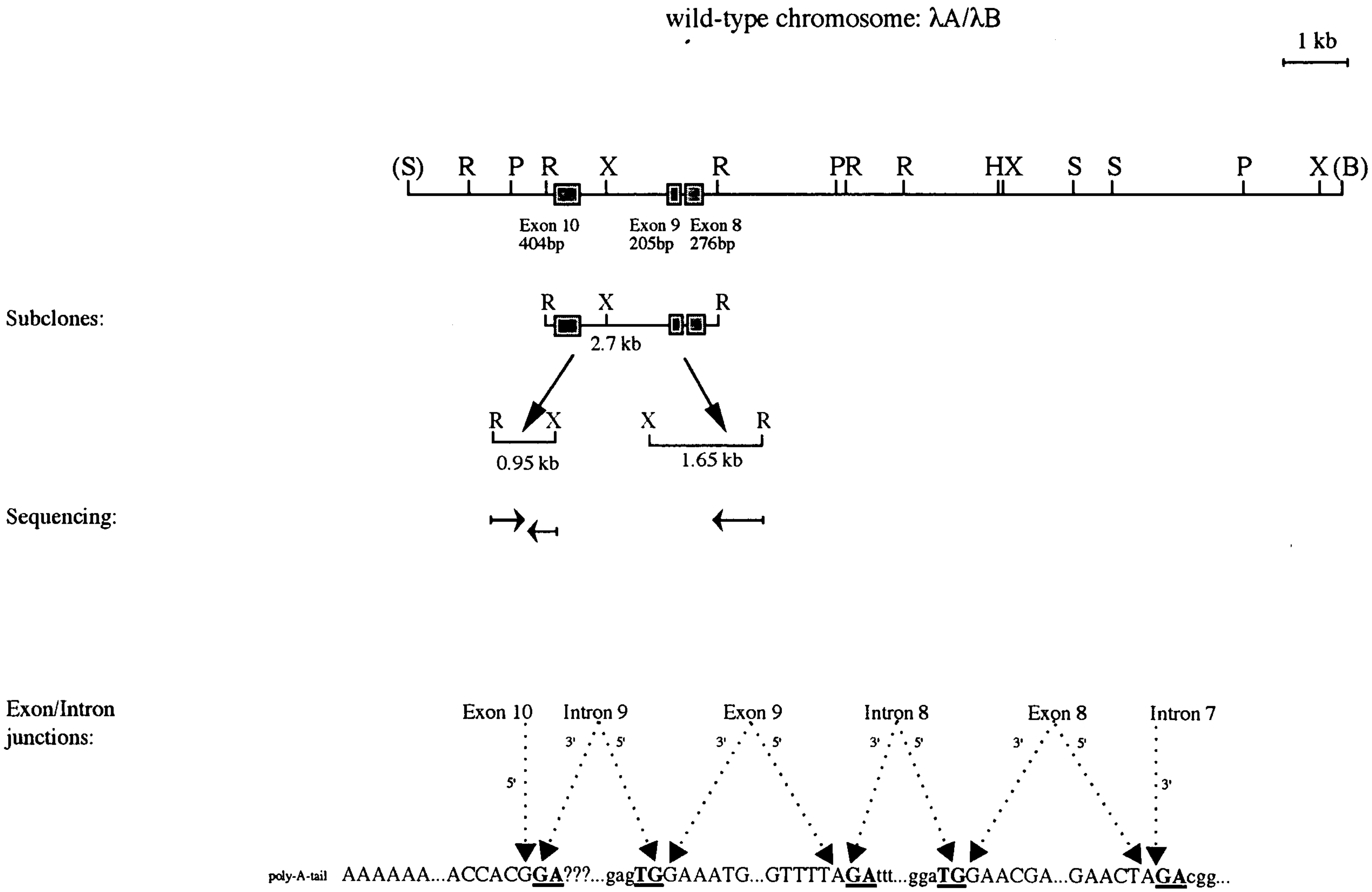
The two *Eco*RI-*Xho*I fragments were subcloned into pBluescriptII (KS-) and partially sequenced. For the sequencing of the 1.65 kb *Eco*RI fragment, the following primers were utilised: M13, reverse M13, SK, T3, 824 and also 916, which was derived from the cDNA sequence (20mer, from position 2031 to 2049 of the cDNA).

Sequencing of exon 8 (276 bp) was aided by the 824 primer which also helped in the identification of the 60 bp intron 8 and 51 bp of the next exon, exon 9. Primer

Figure 15: λA/B map

This figure illustrates λA/B with the following restriction sites: *Eco*RI (R), *Xho*I (X), *Hind*III (H), *Sal*I (S) and *Pst*I (P). The restriction sites in brackets are part of the polylinker of the lambda vector used.

λA/B contains exons 8, 9 and 10 which are all present in a 2.7 kb *Eco*RI–*Eco*RI fragment which was isolated and further subcloned into a 0.95 kb *Eco*RI–*Xho*I fragment and a 1.65 kb *Xho*I–*Eco*RI fragment. Sequencing was carried out as indicated by the arrows and all the exon/intron sequences were established, apart from the intron9/exon10 junction (shown by question marks).


wild-type chromosome: λA/λB
1 kb
(S)
R
P
R
X
R
PR
R
HX
S
S
P
X
(B)
Exon 10
404bp
Exon 9
205bp
Exon 8
276bp
Subclones:
R
X
R
2.7 kb
R
X
X
R
0.95 kb
1.65 kb
Sequencing:
Exon/Intron
junctions:
Exon 10
Intron 9
Exon 9
Intron 8
Exon 8
Intron 7
5'
3'
5'
3'
5'
3'
5'
3'
5'
3'
poly-A-tail
AAAAAA...ACCACGGA???...gagTGGAAATG...GTTTTAGAttt...ggaTGGAACGA...GAACTAGAcgg...

916 was necessary for obtaining the remaining exon 9 sequence of 205 bp in total and its exon/intron junction (fig.15).

The last exon, exon 10 (404 bp), was contained in the smaller *Xho*I-*Eco*RI fragment (0.95 kb). The intron 9/exon 10 junction was not established, since the cDNA primer 709 (18mer, position 2240 to 2257) was used in the 5' to 3' direction which started 42 bp into exon 10. The 3' end of exon 10 was located 94 bp away from the 3' end of the *Xho*I-*Eco*RI fragment.

Figure 16 shows the summary of the organisation of the gene, including all the λ clones and exons isolated.

4.2.3 Summary of sizes and cDNA positions of exons

exon no.	size of exon in basepairs	cDNA position of exons
exon1	466	1- 466
exon2	590	467–1056
exon3	138	1057–1194
exon4	183	1195–1377
exon5	64	1378–1441
exon6	182	1442–1623
exon7	93	1624–1716
exon8	276	1717–1992
exon9	205	1993–2197
exon10	404	2198–2601

All intron/exon junctions sequenced match the consensus for eukaryotes (Mount 1982). Only for exon 4 and 10 the intron/exon junctions were not established.

Figure 16: Map of genomic organisation

This figure shows a summary of all the lambda clones isolated and illustrated in the previous figures. The horizontal line represents approximately 74 kb of DNA containing the *smooth* gene. The line is dotted where an overlap between the lambda clones had not been established, since the precise position of λA22 relative to λSP17 and λRX19 is not known.

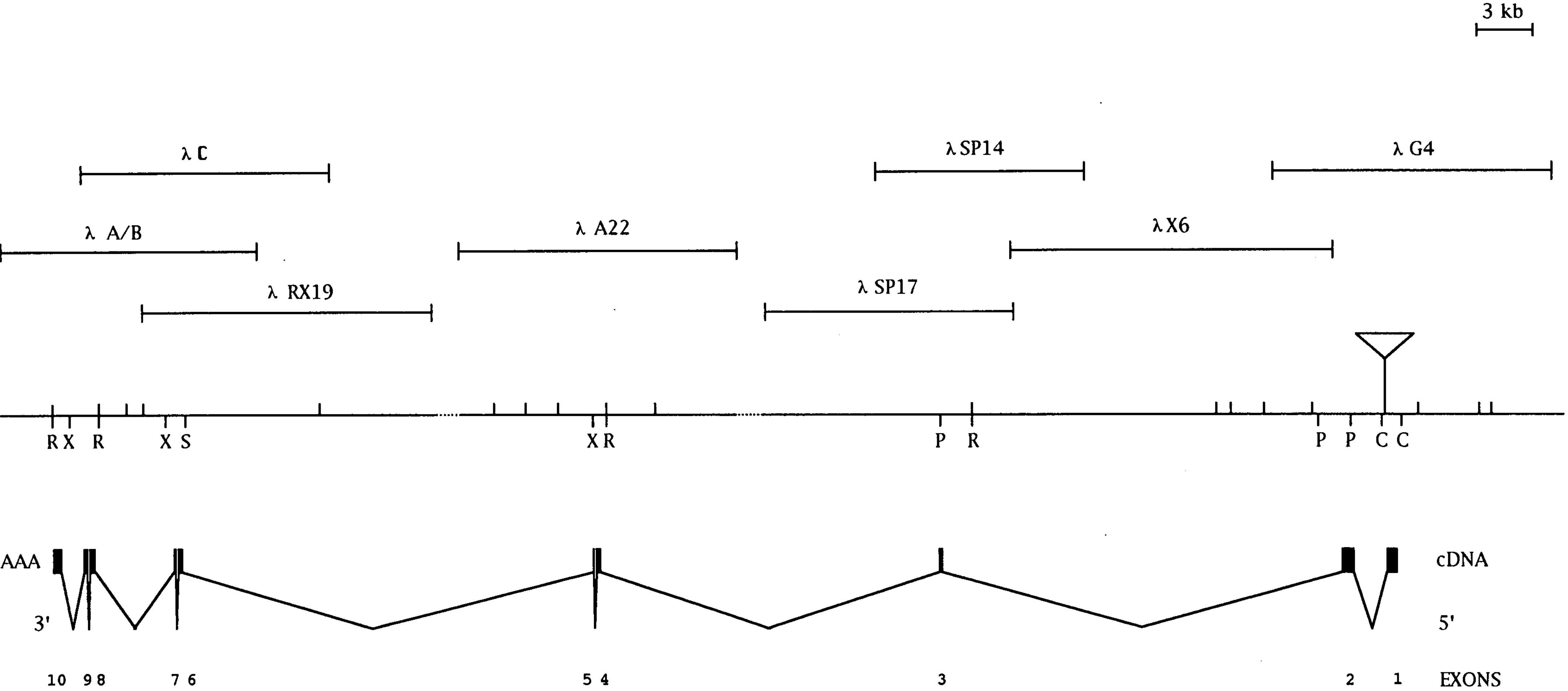
*Eco*RI sites are indicated by ticks above the horizontal line; below it, only those restriction sites are shown which were used to subclone fragments for sequencing, since they contained the exons.

The cDNA is illustrated by black boxes, each representing one exon. The exons are labelled from 1 to 10.

The site of P element insertion into the LP chromosome is indicated by the triangle.

Opposite page 103

Map of lambda clones



4.2.4 Northern blots and developmental expression

In order to compare the length of the cDNA obtained by sequencing with the size of the actual mRNA, Northern blot analysis was carried out (see Methods).

A large fragment of the cDNA was used to probe Northern blots. An internal primer, 575 (nucleotide residues 396→416 of the cDNA), and forward M13 or reverse M13 were utilised, depending on whether the cDNA had been inserted into pTZ18 or 19R, to amplify up a 2226 bp fragment. The PCR product was random-primer-labelled with ^{32}P and used to hybridise to Northern blots of mRNA from different developmental stages.

The photograph (fig.17) shows the pattern of expression of transcripts on a Northern blot of egg, larval, pupal and adult mRNA of the wild-type Samarkand strain, probed with the cDNA probe. As an estimate for the amount of mRNA loaded and in order to obtain a relative size comparison, the filter was subsequently reprobed with part of the α-tubulin 1 gene (Kalfayan *et al.* 1982 and Theurkauf *et al.* 1986), which produces a 1.8 kb transcript constitutively expressed during development.

Single transcripts can be observed in each lane, apart from the egg lane, which shows a smear (which might represent transcripts of smaller sizes, but further evidence is missing). Although the precise size of the hybridisation product was not established, it was found, as expected, to be larger than the α-tubulin 1 transcript. This result is in good agreement with the size of the cDNA obtained by sequencing, 2.613 kb, and suggests that the *smooth* sequence could correspond to the transcript recovered.

While the amount of mRNA loaded for the egg and larval sample is approximately equal, the pupal sample by comparison is vastly overloaded, whereas too little adult mRNA was applied to the gel. A general trend in the pattern of expression of the *smooth* gene can be observed.

Figure 17: Developmental Northern blot

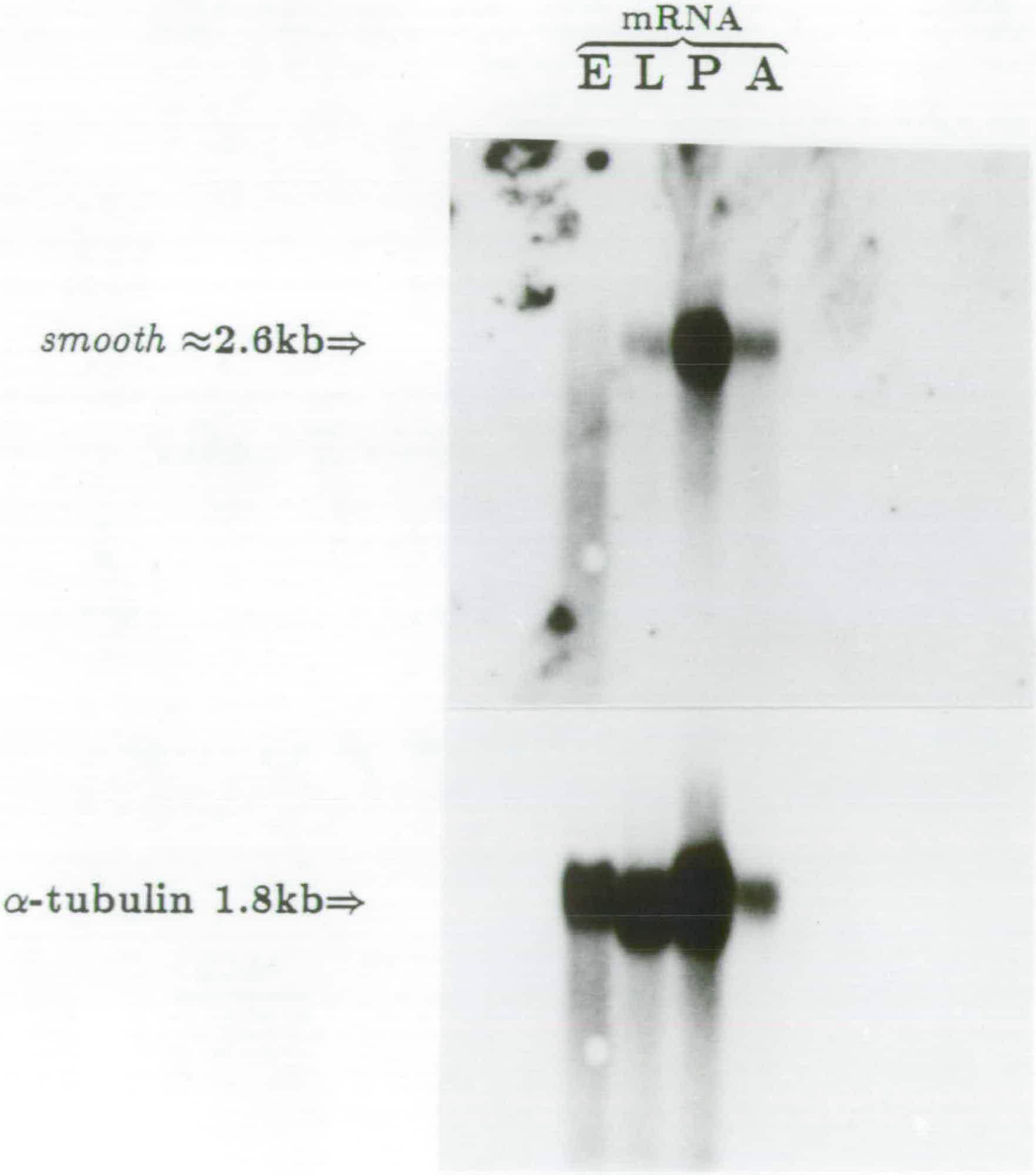
Figure 17 shows the photograph of a developmental Northern blot. The four tracks contain approximately 10μg of egg (E), larval (L), pupal (P) and adult (A) poly(A)$^+$ RNA isolated from the *Drosophila* Samarkand strain. The samples were loaded onto a 1.4% formaldehyde containing agarose gel.

The blot was probed with an almost complete cDNA clone (2226 bp) ^{32}P radiolabelled probe.

After an exposure of 10 days, the blot was reprobed, without prior removal of the cDNA label, with a 1.8 kb α-tubulin ^{32}P labelled probe which served as a loading control.

Opposite page 105

Northern Blot



Varying levels of transcript are detected. Whereas no transcript of comparable size to what is seen in the larvae, pupae and adult is present in the egg, i.e. during the embryonic stages, low levels of hybridisation are first appearing in the larvae. An increased level of expression can be observed in the pupae, but the level of expression is decreased again in the adult. Due to the uneven loading of the gel shown in figure 17, the level of expression in the adult might be approximately equivalent to that of the pupal level. Although there is no evidence suggesting that the *smooth* locus encodes more than one mRNA, there is still the possibility that more than one transcript of similar size can be produced. Transcripts of the same size could come about by alternative splicing and would not be distinguishable on the gel.

4.3 Analysis of the cDNA clone sequence

4.3.1 The putative transcript

It has not been determined whether the cDNA clone isolated represents the full-length transcript, including the 5' and 3' termini. A primer extension experiment, which had not been carried out, would be necessary to verify whether the 5' end of the cDNA clone was present.

A curious phenomenon, noticed at the 5' end of the cDNA clone suggests, that this may indeed be the 5' terminus. An insertion of a single extra G nucleotide residue was observed to be present in the cDNA, but was not found in the corresponding genomic sequences. The addition of an extra G residue has been observed in several full-length cDNA clones, isolated from the same cDNA library (Brown and Kafatos 1988) and are thought to be an artifact produced by the reverse transcriptase attempting to transcribe the mRNA cap (Brown *et al.* 1989 and St. Johnston *et al.* 1991). Although no actual survey has been conducted on the appearance of the additional G residue, the fact that it has been reported to exist at the 5'end of several other loci could suggest that the 5' end of the cDNA is complete.

The genomic sequence 294 bp upstream of the cDNA, corresponding to the 5' end of the HincII fragment sequence, was examined for consensus sequences for the initiation of transcription, i.e. TATA and CAAT boxes. Neither a TATA box at the appropriate region of about 30 nucleotides upstream of the 5' end of the cDNA clone, nor a CAAT box 80 nucleotides upstream was found. The only TATA sequences present were located between nucleotides 9 to 22 (three copies) and 64 to 69 of the HincII fragment. No CAAT box sequence was detected at all. The conclusion can be drawn that either the 5' end of the cDNA is not complete, because at least a TATA and CAAT box is expected, or we are dealing with a different kind of promoter, which does not contain any of these common signals for RNA polymerase II recognition.

Drosophila homeotic genes like *Ultrabithorax*, *engrailed* and *Antennapedia* are examples of genes which do not have the obvious control elements. Transcription in these genes might be controlled by sequences immediately surrounding the transcription start site (Smale and Baltimore 1989). Further investigation would be necessary and hence, the sequences which are involved in the initiation of transcription of the gene remain presently unknown.

The 3' end of the cDNA consists of 12 A nucleotide residues, which should normally correspond to the polyadenylation site. But these 12 A residues do not necessarily represent the poly A-tail. The cDNA library was constructed (Brown *et al.* 1988), using 12 A nucleotides as a first strand primer during the directional cloning process and these were therefore incorporated into the plasmid vector. This has the disadvantage, as pointed out in Yang *et al.* (1991) and St Johnston *et al.* (1991), that truncated 3' ends of cDNAs can be produced, if there is a natural run of A's internal to the cDNA.

When the 3' end of exon 10 was examined at the junction with the genomic DNA, an A rich region was discovered. But since this is the only cDNA isolated for the gene, it is not known whether the 3' end is complete or not.

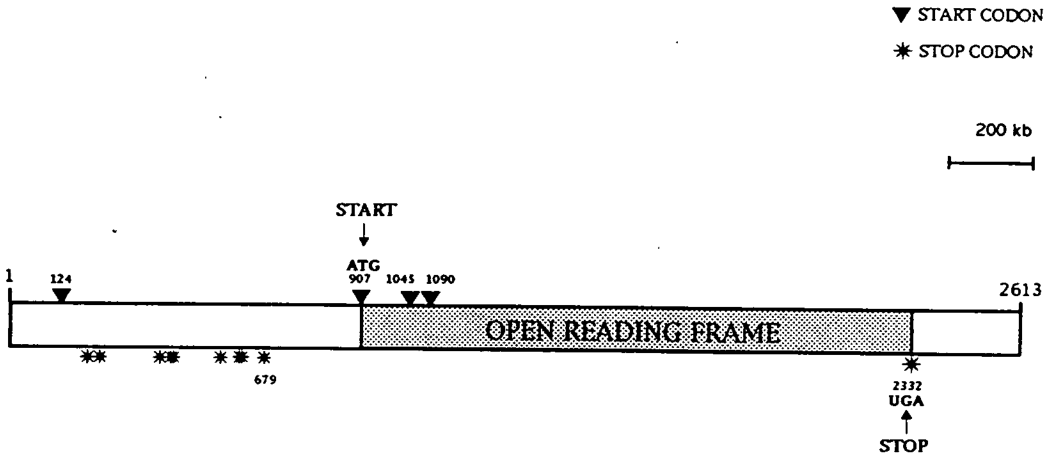
One feature suggesting that the 3' end of the transcript is not present, is the fact that the polyadenylation site is not preceded by a sequence matching the consensus polyadenylation signal: **AATAAA** (Proudfoot and Brownlee 1976). This sequence is highly conserved and can normally be found in a region 11 to 30 nucleotides upstream of the polyadenylation site. The only sequence similar to the A_2TA_3 can be detected at nucleotide residue 2556 with AATTAAA. Further screening of cDNA libraries using the 3' end of the present cDNA clone as a probe, would be necessary to confirm the actual 3' end.

4.3.2 Translation

For the initial analysis of the sequence, the UWGCG FRAMES program (Devereux *et al.* 1984) was used. It draws the six reading frames, showing all start and stop codons of a sequence (see Appendix). Due to the fact that the cDNA library was constructed using directional cloning, only the first three 5' to 3' reading frames were of relevance. All the bars above the line indicate start codons, i.e. methionines, and bars below the line correspond to stop codons.

A single long open reading frame is present in the cDNA sequence, bounded by termination codons at 690 and 2332. This contains a potential initial methionine at position 907.

Figure 18: Sketch of one open reading frame showing a number of start and stop codons



The consensus translational initiation sequence, **C/A A A C/A A U G** (Cavener 1987) is in good agreement with the sequence of the predicted start codon at 907: **C A C A A U G**. Two in-frame ATGs further downstream (at nucleotides 1045 and 1090, respectively) which could also act as the starting methionines, do not match the translational consensus start sequence as well. Therefore the ATG at position 907 is assigned as the initiating methionine.

As the consequence of the start codon being located at position 907, an unusually long untranslated 5' leader sequence is predicted, which again suggests that the 5' terminus of the cDNA clone could be full length.

A long 5' leader sequence with methionine codons upstream from the actual start sequence can produce false translation starts and can therefore cause inefficient translation. Such upstream methionine codons have been proposed as important elements in the translational control of gene expression (Hunt 1985).

Translation would be terminated by the opal stop codon at 2332 (UGA), followed by a 281 bp 3' untranslated region.

4.4 The peptide sequence

The sequence of the long open reading frame (ORF), together with that of its deduced translation product, is shown in the Appendix. The ORF, starting the translation at nucleotide residue 907 and ending with the termination codon UGA at 2332, encodes a putative polypeptide of 475 amino acids with a molecular mass of 51,913 D, as calculated by the UWGCG program PEPTIDESORT (Devereux *et al.* 1984). The isoelectric point of the protein was predicted to be 8.74 – a highly basic protein.

The codon preference data (as calculated by the CODONFREQUENCY and CODONPREFERENCE programs of UWGCG) obtained for the putative protein are in good agreement with the data available for *Drosophila* (M. Ashburner

1989) in showing a typical *Drosophila* codon bias throughout the length of the ORF. The amino acid composition of the protein obtained does not diverge much from the average in *Drosophila*, as compiled by M. Ashburner (listed in Smoller *et al.* 1990).

Inspecting the primary structure of the polypeptide sequence, there are two notable features. The amino terminal of the protein is very glycine (G) and glutamine (Q) rich. Although the percentage of glycine constitutes 8.2% of the whole protein and does not diverge much from the average found in *Drosophila* (7.2%), the percentage of glycine in the first 60 amino acid residues is increased to 18.3%.

Similarly, the overall percentage of glutamine in the smooth protein of 4.4% represents an even lower frequency than the average in *Drosophila* (5.0%). On the other, hand the glutamine residues make up 13.3% of the first 60 amino acids at the amino terminus.

The role of these relatively high frequencies of glycine and glutamine are not known, but it has been proposed that regions with a low charge and a small frequency of hydrophobic residues have an effect on the tertiary structure by forming coils which provide flexibility to the protein (Brendel and Karlin 1989).

The other point of interest is the higher than average frequency of proline residues and their distribution. Due to the cyclic structure of the proline residue, a large number of prolines causes an increased number of turns or coils which again affects tertiary structure of the protein. The average percentage of proline residues in *Drosophila* has been calculated as 5.7%, whereas the smooth polypeptide shows a percentage of 7.6%. Apart from two major gaps between amino acid position 106→157 and 235→324, the proline residues are relatively evenly distributed over the protein. Although three possible glycosylation sites are present, this does not necessarily imply that glycosylation occurs at those positions.

In vitro transcription of the cDNA, followed by *in vitro* translation of the transcription product (see Methods), synthesised a protein with a mobility of about 54 kD on a SDS-polyacrylamide gel (photograph fig.19). This approximately corresponds to the molecular weight of 51.9 calculated.

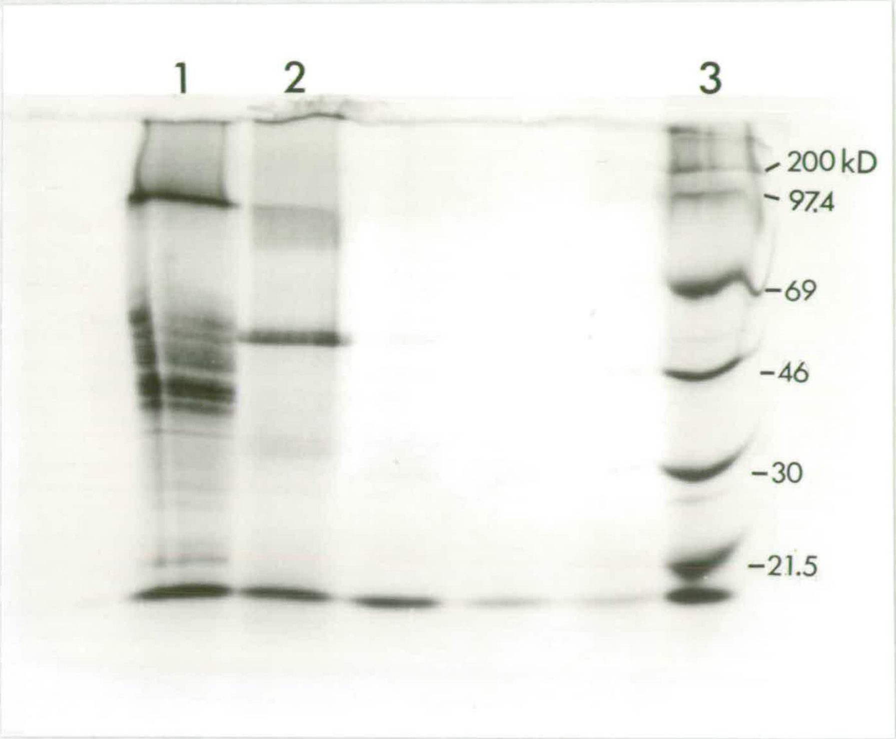


Figure 19: Photograph of the translation product on a SDS-polyacrylamide gel. Track 1: Brome Mosaic Virus control RNA. Track 2: translation product of smooth transcript. Track 3: Molecular weight size markers.

4.5 Protein homologues

A computer search in the GenEMBL database using the TFASTA program (Pearson and Lipman 1988), which searches the database for protein sequences similar to the one in question, revealed a number of proteins showing a remarkably high level of sequence homology to the isolated protein. The protein with the highest identity score listed by the program was encoded by the human mRNA for heterogeneous nuclear ribonucleoprotein L; then followed by a group of eight pyrimidine tract-binding (PTB) proteins, including the heterogeneous nuclear ribonucleoprotein I, which has separately been identified as both a PTB and a hnRNP protein (i.e. the human hnRNP I is identical in its protein sequence to the human pyrimidine-tract binding protein isoform four, PTB-4).

The nine top scoring proteins are listed overleaf:

Protein	% of identity	length of overlap	total length
human heterogeneous nuclear RNP L protein (Piñol-Roma *et al.* 1989)	59.7 %	119 a.a.	2033 bp
rat pyrimidine binding protein 1 (PYBP 1) (Brunel *et al.* 1991)	30.1 %	342 a.a.	2723 bp
rat pyrimidine binding protein 2 (PYBP 2) (Brunel *et al.* 1991)	27.1 %	328 a.a.	2208 bp
H. sapiens PTB-2 gene (Patton *et al.* 1991)	30.3 %	356 a.a.	3188 bp
H. sapiens PTB-1 gene (Patton *et al.* 1991)	30.4 %	342 a.a.	3131 bp
H. sapiens PTB-4 gene (Patton *et al.* 1991)	24.4 %	270 a.a	3209 bp
human polypyrimidine tract-binding protein (Gil *et al.* 1991)	30.4 %	342 a.a.	3090 bp
human heterogeneous nuclear RNP I protein (Ghetti *et al.* 1992)	24.4 %	270 a.a.	3319 bp
mouse 25 kDa nuclear protein (PTB) (Bothwell *et al.* 1991)	29.2 %	342 a.a.	3083 bp

Table 4.1: Smooth protein comparisons using the TFASTA program

Pairwise comparisons of the overall similarity of the protein and the cDNA (including the ORF only) sequences, using the GAP program of UWGCG, established the following similarities and identities (listed in the same order as the TFASTA protein sequence comparison output, apart from the cDNA comparison):

Sequences compared	% of similarity	% of identity
smooth protein x hnRNP L protein	64.9 %	44.1 %
smooth protein x rat PYBP 1	52.9 %	29.1 %
smooth protein x rat PYBP 2[1]	56.0 %	32.0 %
smooth protein x hum PTB-2 protein	53.6 %	30.3 %
smooth protein x hum PTB-1 protein	52.3 %	29.1 %
smooth protein x hum PTB-4 protein	54.8 %	31.3 %
smooth protein x human PTB protein	52.3 %	29.1 %
smooth protein x hnRNP I protein	54.8 %	31.3 %
smooth protein x mouse PTB protein	52.4 %	28.6 %
cDNA smooth seq x cDNA hnRNP L seq	52.2 %	52.2 %

Table 4.2: GAP sequence comparisons of whole length sequences

[1] only partial sequence available in database

The GAP program determines percentages of similarity and identity for two aligned sequences (DNA or peptide sequences). An optimal alignment is achieved when the number of identical bases or amino acids is the largest and the number of gaps is the smallest.

For DNA sequence comparisons there is no difference between the similarity and the identity score, since the four nucleotides are being treated as different entities and therefore are either identical, or not identical, but never similar. The similarity score can therefore be ignored.

The matter becomes more complicated with peptide sequences, where up to 22 "different" amino acids have to be compared and aligned. In this case, distinctions can be made between identical and similar amino acids, where identical means exactly the same amino acid. The degree of similarity on the other hand is based on the evolutionary distance between the two amino acids under comparison. The GAP program deals with this by using a symbol comparison table which contains protein weight matrices according to the scheme proposed by Dayhoff *et al.* (1978). It assigns appropriate scores to each pair of amino acids, with a fixed score for identity and a range of scores, dependent on the degree of relatedness to the similar amino acids. The similarity scores are normally higher than the identity scores.

In the table (4.2) above, the similarity and identity scores are listed. The TFASTA program only considers a single region of overlap of a sequence producing the highest percentage of identity (it does not list the similarity score). The GAP program, on the other hand, compares sequences over their whole length. This explains the differences in the identity score obtained for the GAP program and the TFASTA program. Where, for example, the TFASTA program produced an identity score of 59.7% between the hnRNP L protein and the smooth protein, the GAP program showed a reduced identity score of 44.1%. The TFASTA comparison only considered a sequence of overlap of 119 amino acids, generating the highest percentage of identity, but ignored the rest of the sequence, which contained other regions of strong homology.

As a comparison, a database search using the smooth cDNA nucleotide sequence was carried out using the FASTA program (table 4.3). The heterogeneous nuclear RNP L protein was still detected. It again gave the highest identity, but was followed by four other DNA sequences which did not produce any similarity over a great length.

It is notable that none of the other proteins which show similarity to the putative smooth protein, apart from the hnRNP L, are present in the compiled table of the most similar DNA sequences.

The table below shows a list of the five highest scoring DNA sequences:

DNA sequence of the gene	% of identity	overlap	length
1 human heterogeneous nuclear RNP L protein	62.7 %	410 bp	2033 bp
2 Kluyveromyces lactis GAL 11 gene	63.7 %	91 bp	4159 bp
3 D. discoideum protein kinase gene	56.6 %	258 bp	2696 bp
4 D. discoideum alpha-mannosidase gene	72.0 %	50 bp	4615 bp
5 D. discoideum phosphodiesterase gene	63.1 %	84 bp	6372 bp

References in order 1→5: Piñol-Roma *et al.* 1989; Dickson *et al.* 1991; Buerki *et al.* 1991; Schatzle *et al.* 1989; Podgorski *et al.* 1989.

Table 4.3: Result of FASTA (DNA sequence comparison)

With regard to the high level of similarity between the human hnRNP L protein and the smooth protein, it is considered that smooth represents the *Drosophila* homologue of the human hnRNP L protein.

Comparison between smooth and hnRNP L

Comparison at the DNA level:

To illustrate the large degree of homology between the *smooth* gene and the human hnRNP L DNA sequence, the sequences were run through the COMPARE program and subsequently the DOTPLOT program of UWGCG. Figure 20 shows a DOTPLOT comparison between the smooth cDNA sequence and the human hnRNP L protein cDNA sequence.

The significant regions of homology are depicted as short stretches of diagonal lines between the two coordinates, with the horizontal line representing the cDNA smooth sequence and the vertical line the cDNA of the hnRNP L protein. A continuous diagonal line would indicate complete homology between two sequences. The four major regions of almost identical sequences are the following:

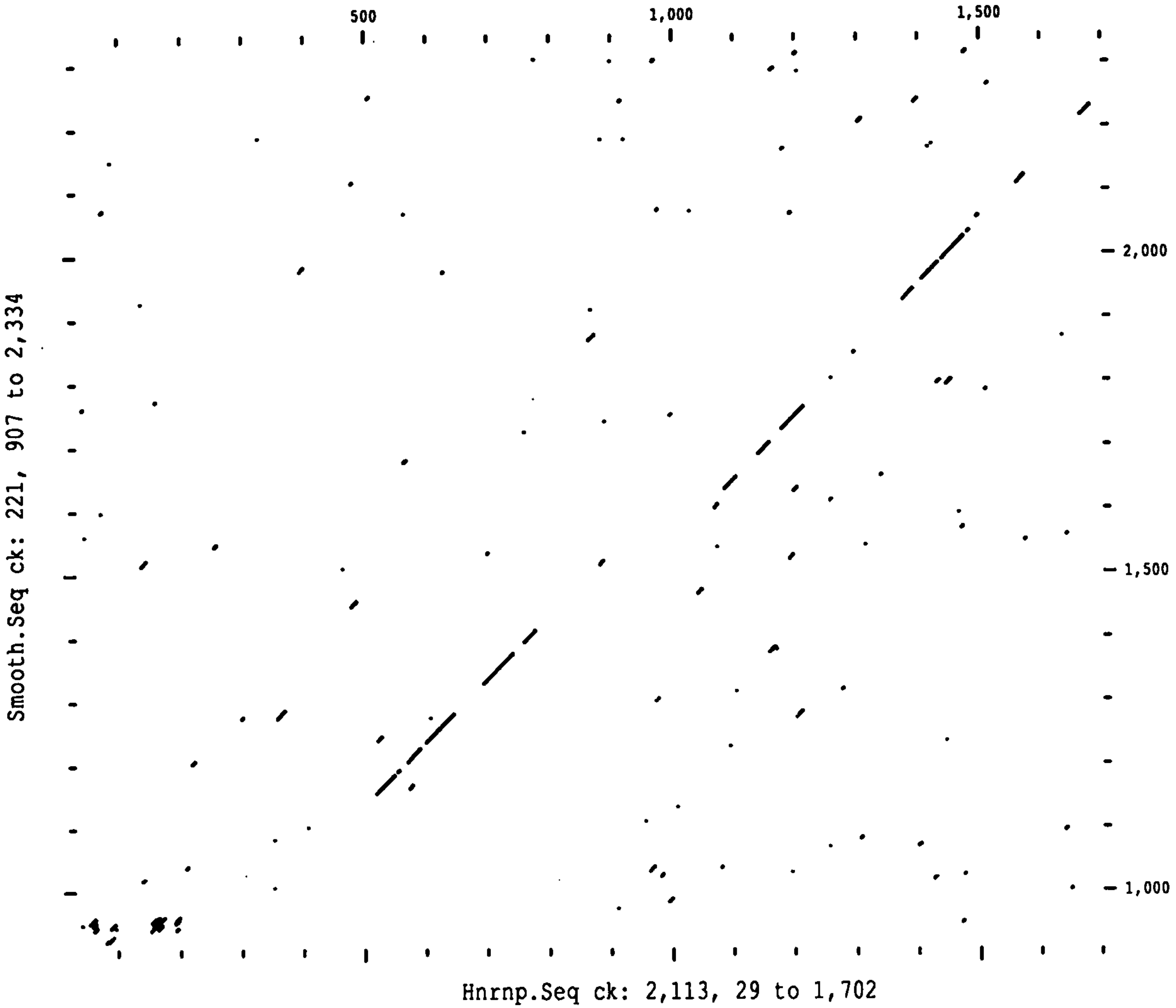
Figure 20: DOTPLOT of smooth cDNA and the human hnRNP L cDNA

Figure 20 represents a dot matrix comparison (DOTPLOT) between the hnRNP L DNA sequence, with only the coding sequence (from 29 bp to 1702) depicted along the horizontal axis, and the smooth DNA sequence, again only containing the coding sequence (from 907 bp to 2334 bp) along the vertical axis.

The DOTPLOT was produced with the UWGCG package by first of all establishing a comparison table, using the COMPARE program with the window setting of 21 and a stringency of 14.0.

The regions of highest homology between the two sequences can be seen as short diagonal lines.

Opposite page 118


500
1,000
1,500
2,000
1,500
1,000
Smooth.Seq ck: 221, 907 to 2,334
Hnrnp.Seq ck: 2,113, 29 to 1,702

	bp position in smooth sequence	bp position in hnRNP L sequence
1	1151→1287	516→ 651
2	1320→1425	686→ 780
3	1679→1768	1132→1220
4	1905→2051	1351→1497

Table 4.4: Four regions of significant homology between the smooth cDNA sequence and the hnRNP L protein cDNA sequence

A number of smaller regions of homology can also be detected along that diagonal line.

Comparison at the protein level:

Carrying out the same analysis for the protein sequence between the putative smooth protein and hnRNP L protein, an even higher degree of homology can be observed (fig.21) in the protein DOTPLOT comparison. Starting at the amino acid position 80 in the smooth protein and 160 of the hnRNP L protein, the homologous region extends to amino acid position 175 and 254 respectively. After a small gap, a renewed region of high similarity is produced between 237 of smooth which corresponds to 347 of the hnRNP L protein. This diagonal line is almost continuous until amino acid position 300 in the smooth and 410 in the hnRNP L protein. The last region of homology starts at position 328 in smooth and 435 in the hnRNP L protein and extends with only two brief stretches of weaker homologies to position 440 and 552, respectively.

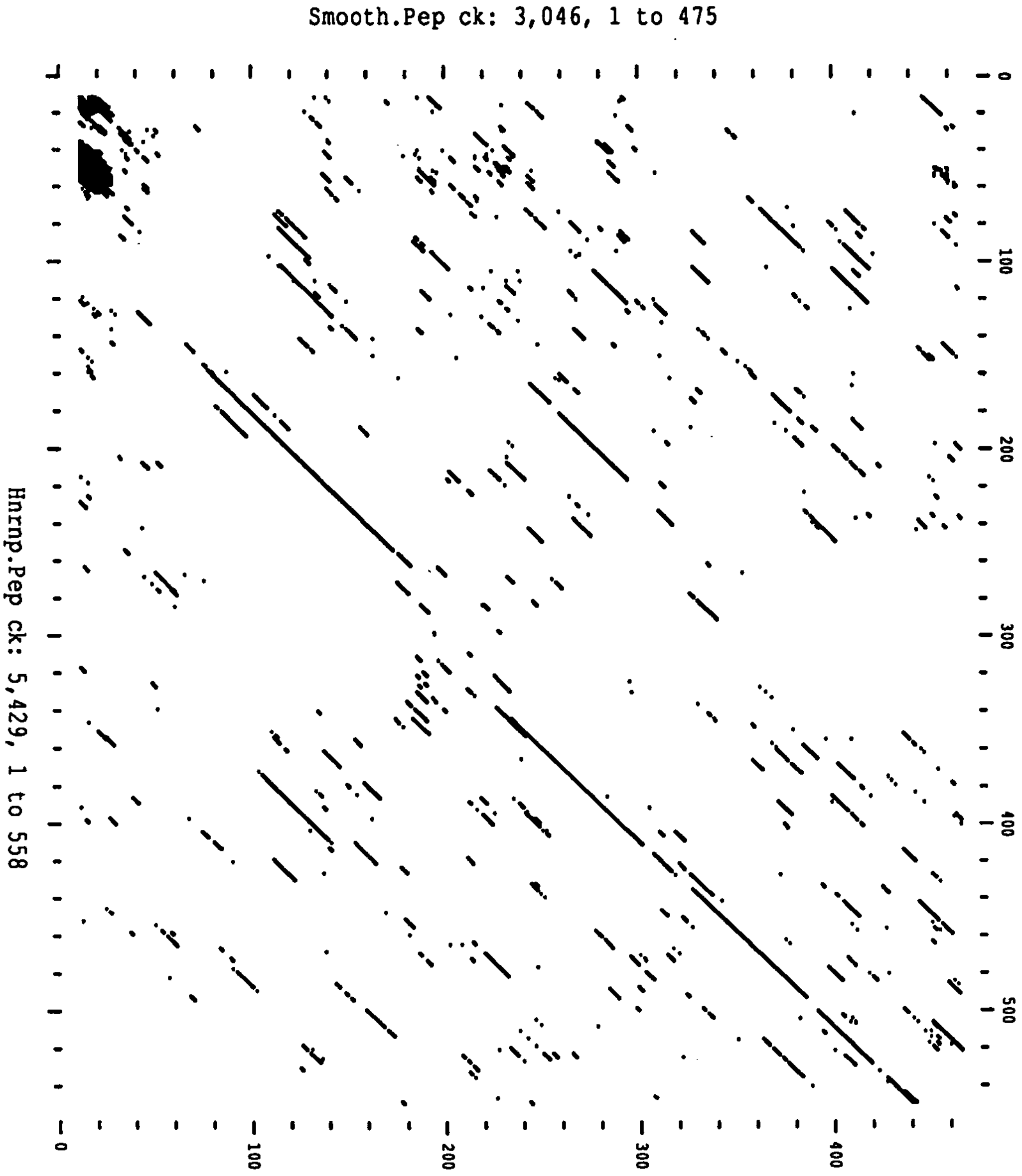
A direct protein sequence comparison of the three distinct regions of highest homology between hnRNP L protein and the smooth protein are depicted at the end of the next chapter. A complete comparison of the two protein sequences can be found in the Appendix.

To illustrate internal sequence repeats in the smooth protein itself, a DOTPLOT in figure 21a shows a smooth versus smooth comparison.

Figure 21: DOTPLOT of smooth protein and the human hnRNP L protein

Figure 21 shows a DOTPLOT comparison of the complete hnRNP L protein amino acid sequence on the X-axis and the complete smooth amino acid sequence on the Y-axis.

The DOTPLOT was produced by using the DOTPLOT program of the UWGCG package by first establishing a comparison with the COMPARE program using a window setting of 20 and a stringency of 9.0



Smooth.Pep ck: 3,046, 1 to 475
Hnrnp.Pep ck: 5,429, 1 to 558
0
100
200
300
400
0
100
200
300
400
500

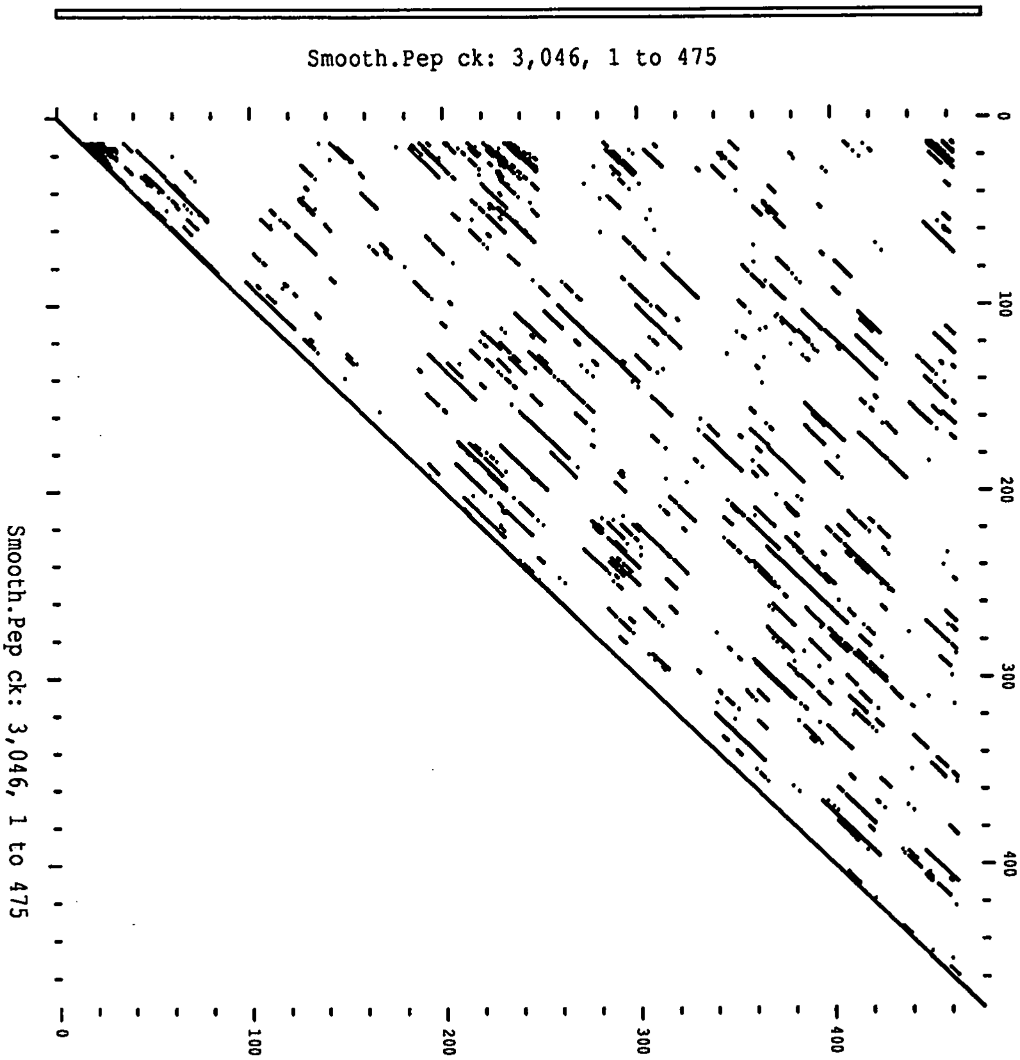
Figure 21a: DOTPLOT of smooth protein versus the smooth protein protein

This figure shows a DOTPLOT comparison of the complete smooth amino acid sequence on the X-axis versus the same sequence on the Y-axis. The window setting was 25 and the stringency 9.0 in the COMPARE program.

Some small areas of sequence repeats can be found.

DOTPLOT of: smooth.pnt;2 Density: 542.05 May 30, 1993 16:26

COMPARE Window: 25 Stringency: 9.0 Points: 6,503



4.5.1 Homology to pyrimidine binding proteins

The other eight top-scoring proteins apart from the human hnRNP L protein, which are listed in the TFASTA table (4.1) and compared by the GAP program (table 4.2), all belong to the same group of pyrimidine binding proteins exhibiting a remarkable degree of homology. The rat PYBP 2 (polypyrimidine binding protein 2) has only been partially entered into the database (the amino terminus seems to be missing), but shows complete homology to the rat PYBP 1, apart from an insertion of 25 amino acids between positions 296 and 297 of PYBP 1.

Different isoforms of the human pyrimidine-tract binding (PTB) proteins are listed, where the human PTB protein is completely homologous in its protein sequence to PTB-1 and also the PTB-4 isoform is equivalent to hnRNP I.

The human and the murine polypyrimidine-tract binding proteins show an identity of 97% to each other at the protein level. The same value of 97% applies to the level of identity found between the rat PYBP 1 and the human polypyrimidine-tract binding protein.

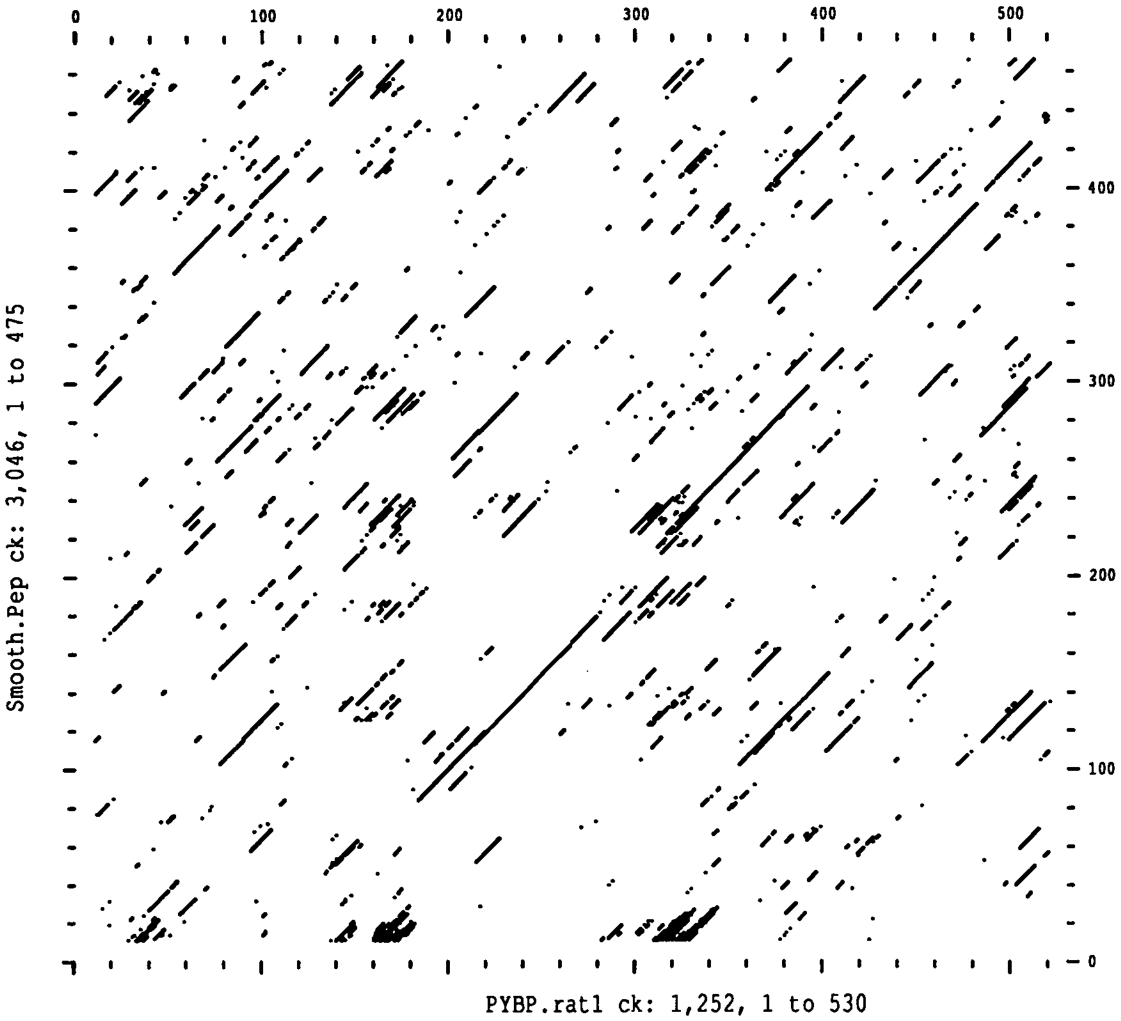
When this group of pyrimidine binding proteins is compared to the putative smooth protein, approximately the same extent of identity is found, not unexpectedly, over a conserved region for all of them. There is no significant homology between the pyrimidine binding proteins and the smooth protein at the DNA level.

The GAP comparison over the complete length of the proteins (table 4.2) again reveals very similar results. The percentage of identity of the pyrimidine binding proteins and the smooth protein is considerably smaller compared to the values obtained for the smooth protein and the hnRNP L protein. On the other hand, the DOTPLOT analysis of the comparison between one of the pyrimidine binding proteins, the rat PYBP 1 (fig.22), and the smooth protein shows extensive similarities over large regions nevertheless.

Figure 22: DOTPLOT of smooth protein and the rat polypyrimidine-tract binding protein

Figure 22 shows a DOTPLOT of the rat polypyrimidine binding protein (PYBP) 1 amino acid sequence along the X-axis and the smooth amino acid sequence along the Y-axis.

The UWGCG package programs COMPARE and DOTPLOT were used and a window setting of 20 and a stringency of 9.0 was used for the comparison.


0
100
200
300
400
500
400
300
200
100
0
Smooth.Pep ck: 3,046, 1 to 475
PYBP.rat1 ck: 1,252, 1 to 530

According to the DOTPLOT output, the first region of similarity starts at amino acid position 84 of the smooth protein, extending to position 175 which corresponds to position 182 to 276 of the rat PYBP 1 protein. The continuous stretch of homology can be found from position 228 to 299 in the smooth protein and 326 to 394 in the PYPB 1. The last major region of similarity is present from position 336 to 390 in the smooth protein and corresponds to amino acid 427 and 482 in PYPB 1.

Table of the amino acid positions of regions of homology common to the human hnRNP L and the smooth protein and the rat pyrimidine binding and the smooth protein:

1		2	
hnRNP L protein	**smooth protein**	**rat PYBP 1**	**smooth protein**
160→254	80→175	182→276	84→175
347→410	237→300	326→394	228→299
435→552	328→440	427→482	336→390

Table 4.5: Regions of protein homology as identified by DOTPLOTS

The above table indicates that there are certain conserved regions in the smooth protein, since the three main areas of homology with the hnRNP L protein and the PYBP 1 cover approximately the same regions. Due to the high percentage of homology between PYBP 1 and the PTB proteins, the DOTPLOTS comparing the human or murine PTB protein with the smooth protein are similar in appearance (not shown).

The hnRNP L and the pyrimidine binding proteins have been associated with RNA binding and also specific single-stranded DNA binding, a property which is common to many of the RNA binding proteins, specifically in the case of the rat PYBP 1 and 2 proteins (Brunel *et al.* 1991) and the murine PTB protein (Bothwell *et al.* 1991).

Further discussion of the proteins homologous to smooth and analysis of their RNA binding domain can be found in the subsequent chapter.

Chapter 5

Proteins homologous to the smooth protein in vertebrates

5.1 Homologous proteins

5.1.1 Heterogeneous nuclear ribonucleoproteins

The heterogeneous nuclear ribonucleoproteins (hnRNPs) comprise at least 20 major polypeptides designated alphabetically from A through to U according to their molecular weight from 34,000 for A1 to 120,000 for U (Piñol-Roma *et al.* 1988). These hnRNP proteins are very abundant and form a complex in the nucleus of many eukaryotes with the RNA polymerase II transcript (also called pre-mRNA, nascent RNA transcript or heterogeneous RNA). They are thought to be actively involved in the post-transcriptional processing of pre-mRNA which includes splicing, storage and transfer, e.g. nucleo-cytoplasmic transport of mRNA (Piñol-Roma and Dreyfuss 1991; 1992).

The isolation of the hnRNP complex and single hnRNPs from a nuclear extract can be carried out in various ways. Originally, the nuclear proteins were isolated by sucrose gradient sedimentation (Samarina *et al.* 1968; Pederson 1974). But in order to ensure the selective isolation of RNA-binding proteins, the ability of RNA to photo-activate and react with any molecule in close proximity is now utilised (Dreyfuss *et al.* 1984a; Choi and Dreyfuss 1984).

UV cross-linking in intact cells is carried out followed by affinity chromatography on oligo(dT) cellulose and RNase treatment. From these isolated complexes monoclonal antibodies to individual hnRNP proteins are raised (Dreyfuss *et al.* 1984b). The advantage of having monoclonal antibodies has greatly facilitated the isolation of hnRNPs due to the application of immunopurification procedures (Dreyfuss *et al.* 1988; Piñol-Roma *et al.* 1988).

A number of major hnRNPs have been analysed individually e.g. hnRNP A1 (Buvoli *et al.* 1990), C (Piñol-Roma *et al.* 1988; Swanson and Dreyfuss 1988a; 1988b), D (Piñol-Roma *et al.* 1988), I (Ghetti *et al.* 1992), K (Matunis *et al.* 1992c), L (Piñol-Roma *et al.* 1989) and U (Kiledjian and Dreyfuss 1992) proteins. It has been established that hnRNPs are capable of binding single-stranded nucleic acids and showing a high binding affinity for different ribohomopolymers, e.g. U or G. The binding of RNA has been shown to occur at a special region of 80 to 90 amino acids, the RNA-binding domain (RBD) (also called RNA recognition motif, RRM). The RBD contains two highly conserved motifs, the consensus sequences RNP 1 and RNP 2 (CS-RNP 1 and CS-RNP 2, see section on RNA-binding domain), and has been identified in many RNA-binding proteins, including most of the hnRNPs.

Recently studies have been carried out on interactions of hnRNP complexes as a whole to pre-mRNA and their role in splicing (Bennett *et al.* 1992a; 1992b). The hnRNPs have been shown to be part of the earliest detectable complex, the H complex, which assembles on pre-mRNA during *in vitro* splicing reactions. The formation of the H complex is the first step towards the formation of spliceosomes which involves the action of other classes of ribonucleoproteins, specifically the small nuclear ribonucleoproteins (snRNPs) (Guthrie and Patterson 1988). Although the hnRNPs are part of the final spliceosomes, they might not necessarily function as splicing factors. It has been indicated (Bennett *et al.* 1992a; 1992b) that they might facilitate or prevent binding of other factors, e.g. splicing factors.

Human heterogeneous ribonucleoprotein L

The hnRNP L protein has a predicted molecular mass of 60,187 D and is of neutral pH (Piñol-Roma *et al.* 1989). The L protein has only got poorly conserved CS-RNP 1 and 2 elements and has therefore been referred to as a "novel" heterogeneous nuclear RNP protein. Formerly, it was assumed that the L protein contained only two segments (from amino acid position 63–147 and 155–241) which showed a weak homology to each other and to other RNA-binding domains (Piñol-Roma *et al.* 1989). But since more information has become available on the secondary and recently on the tertiary structure of the RNA-binding domain, it was then noted that a third putative RNA-binding domain was present in the L protein. The third domain extends from amino acid position 342 to 425 (Kenan *et al.* 1991). A fourth region also showing some conservation and similarities to RNA-binding domains was listed in Ghetti *et al.* (1992) and extends from amino acid 461 to 549.

The precise function of the L protein has not been established, although it has been demonstrated that the L protein binds to poly-A-containing RNA in intact cells (Piñol-Roma *et al.* 1989). For the human hnRNP A1, C and D proteins, putative high affinity binding sites towards the 3' end of introns on pre-mRNAs have been detected (Swanson and Dreyfuss 1988b) but, so far, such RNA-binding specificity has not been established for the hnRNP L protein. It is therefore not known which and how many of the three (or maybe even four) RNA recognition motifs (RRMs) are involved in RNA-binding.

Another feature, which sets the human hnRNP L protein apart from the other hnRNPs discovered, is that a monoclonal antibody to hnRNP L binds to the giant loops of amphibian lampbrush chromosomes. High concentrations at C-rich transcripts generated at the giant loop were detected. The L protein was the first protein found to become selectively distributed over the nascent nonnucleolar transcripts of lampbrush chromosomes in newts (Piñol-Roma *et al.* 1989).

5.1.2 The pyrimidine tract binding proteins

The group of proteins more closely related to the *smooth* protein than any other heterogeneous ribonucleoprotein, apart from the human hnRNP L protein, are the recently discovered pyrimidine tract binding proteins (PTB).

When searching the database with the rat 1 PYBP, apart from listing the other PTB proteins and the human hnRNP L protein, two *Drosophila* proteins are included in the list of the 25 best scores. The proteins are the elav proteins of *Drosophila virilis* (Yao and White 1991) and *Drosophila melanogaster* (Robinow *et al.* 1988). They exhibit identity scores of 21.6% and 21.0% in overlaps of 278 and 300 amino acids, respectively. The elav protein is a neuronal protein containing RNA-binding consensus squences.

The PTBs also have only poorly conserved RNP 2 and RNP 1 motifs, as had already been found for the human hnRNP L protein. On the other hand, the PTB proteins are still capable of binding RNA and single-stranded DNA, and they do have certain characteristics contained in their RNA binding domain which are important for the binding.

The rat 1 polypyrimidine binding protein, 56,936 D protein, was originally discovered as a transcription factor binding to a polypyrimidine tract present in one of the two strands of DNA regulatory elements, DRI and PRI, of the rat tyrosine aminotransferase (TAT) gene enhancer (Brunel *et al.* 1991). Although PYBP has been isolated from rat liver nuclei as a DNA binding protein, binding to at least three regulatory cis-elements of hepatic genes, it contains four repeats of sequences which exhibit the RNA recognition motifs. This is not an uncommon phenomenon, since other proteins have been discovered sharing an ability to bind to single-stranded DNA as well as RNA, hence belonging to a group of nucleic acid binding proteins. The biological role of PYPB has not been defined (Brunel *et al.* 1991).

The rat 2 protein, which showed the second highest homology, is not being considered any further here, since it is only partially available on the database and seems to be completely homologous to the rat 1 protein apart from a 25 amino acid gap.

The human homologue to the rat 1 PYBP is the pyrimidine tract-binding protein. So far, at least four different isoforms have been isolated from HeLa nuclear extracts. The human PTB proteins were isolated by several groups independently (Garcia-Blanco *et al.* 1989; Wang and Pederson 1990; Gil *et al.* 1991; Patton *et al.* 1991; Ghetti *et al.* 1992).

In an experiment studying proteins involved in binding pre-mRNA, a 62 kD protein from the HeLa nuclear extract was identified to UV cross-link specifically to the pre-mRNA of Adenovirus (Ad10) (Garcia-Blanco *et al.* 1989).

Wang and Pederson (1990) detected an approximately 62 kD protein in the HeLa nuclear extract, capable of cross-linking to a 61 nucleotide RNA fragment which is present in the first intron of the human β-globin pre-mRNA. The 61 nucleotide RNA fragment contains a polypyrimidine tract.

A third group (Mullen *et al.* 1991 and Patton *et al.* 1991) who were working on the identification of splicing factors involved in the mutually exclusive exon selection in rat α-tropomyosin, came across a 57.2 kD protein, which specifically bound to polypyrimidine tracts of introns.

The purified PTB protein can be resolved as a doublet (and in some cases as a triplet) of bands, as visualised on SDS-polyacrylamide gels after silver-staining (Garcia-Blanco 1989, Gil *et al.* 1991 and Patton *et al.* 1991). Sequence analysis of the PBT proteins revealed that the 62 kD and the 57.2 kD proteins are identical (Gil *et al.* 1991 and Patton *et al.* 1991).

Gil *et al.* (1991) described three isoforms: the prototype PTB (also called PTB-1) with 531 amino acids and a molecular mass of 57,220 D, PTB-2, which consists of 550 amino acids with an insertion of 57 nucleotides at position 921 of the prototype

PTB nucleotide sequence (maintaining the reading frame) and PTB-3 which has only been reported by Gil *et al.* (1991) and is due to an addition of 77 nucleotides at position 921 of PTB. The reading frame is shifted, resulting in a truncated 42.8 kD protein.

The PTB-1 protein has been identified as being a subunit of a large complex necessary for splicing (Patton *et al.* 1991). The human PTB protein co-purifies with a 100 kD (now called PTB-associated splicing factor, PSF (Patton *et al.* (1993)) and a 33 kD protein, where the 100 kD protein has been found to play an important role as an essential splicing factor during the splicing process in association with the PTB protein. The human PTB protein has been shown to bind to pre-mRNAs, specifically to the polypyrimidine tract of introns (Gracia-Blanco *et al.* 1989), which are usually located immediately upstream of the 3' splice junction and downstream of the branch site.

A series of deletions at the 3' end of the first intron of the AdML (Adenovirus major late transcription unit) helped to identify a region critical for the binding of PTB proteins (Garcia-Blanco *et al.* 1989). Subsequently, *in vitro* mutation experiments (Gracia-Blanco *et al.* 1989) indicated that PTB proteins might be involved in the early stages of splice recognition, pre-spliceosome assembly and remain associated to pre-mRNA during the formation of spliceosomes. The human PTB protein is thought to facilitate or enhance the binding of the U2 snRNP, which is involved in the recognition of the branchpoint and 3'-splice-site region of the pre-mRNA. The experiments revealed that, when the 3' splice site (AG) was intact and the PTB protein bound to the polypyrimidine tract, the U2 snRNP complex was formed, resulting in efficient splicing. On the other hand, when the binding of the PTB protein was prevented (e.g. due to a deleted polypyrimidine tract in the intron), formation of the U2 snRNP complex was reduced. The fact that the absence of PTB protein only reduces the formation of the U2RNP complex indicates that a number of other factors might be involved in this process. One of these is the U2 auxiliary factor (U2AF), which was also found to bind specifically the polypyrimidine tract of introns (Zamore *et al.* 1992).

The murine homologue of the human PTB was isolated from murine plasmacytoma nuclear extracts (Bothwell *et al.* 1991), as part of a complex consisting of three nucleic acid binding proteins (100 kD, 35 kD and 25 kD). The 25 kD protein has been identified to be a proteolytic product derived from the 56,782 D murine PTB protein. The 25 kD protein contains an RNA-binding domain which has been shown as being sufficient for protein binding. The murine PTB protein was originally isolated as a DNA binding protein showing specific binding to single stranded nucleic acid (Ballard *et al.* 1988).

5.1.3 hnRNP I protein and the human PTB protein

Very recently, the DNA and polypeptide sequence of the human heterogeneous nuclear RNP I protein was established (Ghetti *et al.* 1992). The most surprising discovery of the predicted 557 amino acid protein with a molecular mass of 59,632 D and a pI of 9.86, is that it represents another isoform of the human PTB protein. hnRNP I is almost completely identical to the PTB-4 protein (Patton *et al.* 1992), apart from a 7 amino acid insertion at position 298 of the protein sequence.

The DNA sequences of hnRNP I and PTB-4 are not entirely identical. First of all, they differ in length, mainly due to the fact that the hnRNP I DNA has a longer leader sequence. When the sequences are aligned, 16 gaps are found in the 3' untranslated region and three third base polymorphisms in the coding sequence. These polymorphisms do not make any difference to the translation product.

Since the hnRNP I and PBT proteins are identical, the characteristics and the function of the proteins must also be the same, i.e. hnRNP I will specifically recognise polypyrimidine tracts and therefore be involved in the 3' splice site selection.

The hnRNP L protein exhibits a significant homology to the hnRNP I protein. The GAP analysis using UWGCG produced a 28.3% identity score and a 53.8% similarity. A DOTPLOT analysis showed that there are four regions of specific similarity between the two proteins, suggesting that these regions of approximately 80 amino acids represent the RNA binding domain.

Likewise, the putative smooth protein is related in a similar way to the hnRNP I protein as the hnRNP L protein is. The GAP comparison even shows a slightly stronger homology to hnRNP I, with 31.3% identity (compared to 28.3% between hnRNP I and L) and 54.6% of similarity (compared to 53.8% between hnRNP I and L). Although four regions of similarity have been described between hnRNP L and hnRNP I (Ghetti et al. 1992), homology between the smooth and the hnRNP I protein is largely confined to three regions (just as was reported for smooth and the hnRNP L protein and smooth and the rat 1 PYBP earlier, see table 2.5).

Ghetti *et al.* (1992) also made some interesting observation regarding the staining pattern of anti hnRNP I antibody. They noticed that the staining pattern of the anti I monoclonal antibody differed from that found for other hnRNPs. Apart from a high concentration of staining in the nucleoplasm, the antibody was also found to stain the perinucleolar structure. The cytoplasmic staining with hnRNP I antibody was also higher compared to that of other hnRNPs, which hardly stain the cytoplasm at all.

5.2 The RNA-binding domain

5.2.1 Description of RNA-binding domains

The amino acid sequence of the putative smooth protein has revealed three distinct regions of homology to the polypyrimidine binding proteins (rat, murine and human) and an even higher overall degree of homology to the hnRNP L protein, but also including three major regions.

These regions of homology bear a distant relationship to the RNA-binding domains (RBD) which have been discovered in species as diverse as *E.coli* and *Homo sapiens*, including *Saccharomyces cerevisiae*, *Zea mays* and *Drosophila melanogaster.* The 80-90 amino acid long sequences, also called RNA recognition motifs (RRM), consist largely of aromatic and hydrophobic residues, were first discovered in the yeast poly-A-binding protein (PABP) which contains four copies of such a domain (Adam *et al.* 1986 and Sachs *et al.* 1986).

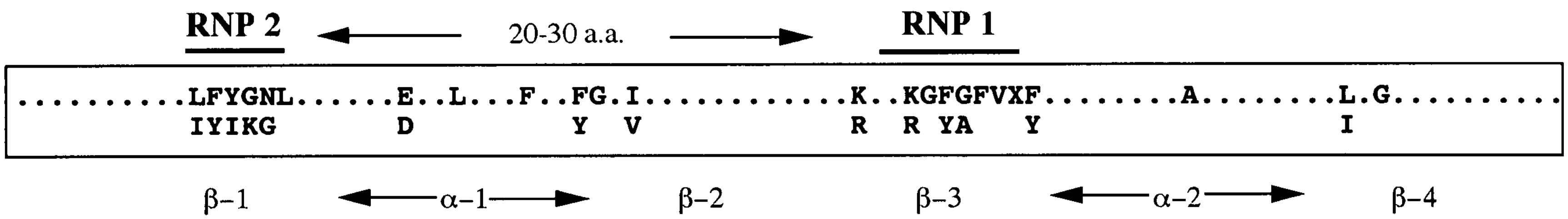
Although the overall sequence similarity varies at the protein level immensely, being only loosely conserved amongst the different nuclear, cytoplasmic and organellar RNA-binding proteins, two RNP consensus sequences (RNP-CS) separated by approximately 30 amino acids are highly conserved. And although there are approximately 21 "conserved" amino acids (conserved in the sense of conserved amino acid type, i.e. aromatic, aliphatic and basic) spread over the extend of the RBD domain, the most conserved oligopeptide regions are the RNP-1 and RNP-2. The RNP-1 is an octamer with the consensus **(R/K)GF(G/A)FVX(F/Y)** located more 3' in the RNA-binding domain than the later discovered and less well conserved hexamer RNP-2 consensus sequence, **LF(V/Y)GNL** (fig.23a).

Many of these proteins with RRMs have been shown to bind single-stranded nucleic acids to differing extends. The type of functions these proteins carry out can be divided into two groups: first of all proteins with basic housekeeping functions,

Figure 23a: Diagram of RNA-binding domain

This figure shows a diagram of a RNA-binding domain. It normally consists of 80-90 amino acids. The RNP 1 and RNP 2 consensus sequences are shown. The most conserved amino acids outside the consensus sequence have been included. The position of the secondary structure pattern, the α helices and the β sheets have been indicated.

RNA binding domain


RNP 2
20-30 a.a.
RNP 1
..........LFYGNL......E..L...F..FG.I.............K..KGFGFVXF........A........L.G.........
IYIKG D Y V R R YA Y I
β–1
α–1
β–2
β–3
α–2
β–4

e.g. snRNPs and hnRNPs and secondly, proteins with developmentally important roles: Sxl, tra-2, elav and MA 16 in maize.

Alignment of RNA-binding domains

Figure 23b illustrates the alignments of the conserved RNA-binding domains of a number of human ($) and Drosophila (#) proteins based on the structure of the human U1 small nuclear ribonucleoprotein A (alignment carried out as in Kenan *et al.* 1991).

Two RNA-binding domains have been identified in the 282 amino acid long sequence (Sillekens *et al.* 1987) of the snRNP U1 A, one of which is located at the amino-terminal end (domain 1 amino acid position 11→91) and the other at the carboxy-terminal end(domain 2 210→282). The structure of the RNA-binding domain 1 of the U1 small nuclear ribonucleoprotein A (snRNP U1 A) has been under intense investigation and analysis of the crystal structure by X-ray crystallography (Nagai *et al.* 1990 and Jessen *et al.* 1991) as well as analysis of the global folding by NMR spectroscopy (Hoffman *et al.* 1991) has been carried out.

So far, only the first domain has been shown to bind to RNA, specifically to stemloop II of snRNA U1 (Scherly *et al.* 1989 and 1990; Jessen *et al.* 1991). The second domain (not included in fig. 23b) does not contain a well conserved RNP-1 consensus and it has been proposed that domain 2 might fulfil a role in binding proteins involved in the splicing complex, but whether this is indeed the case is not known (Scherly *et al.* 1989).

The figure has been divided up into two sections, with A representing RNA-binding domains with well conserved RNP-1 and RNP-2 consensus sequences and B listing RNA-binding domains of proteins without conserved RNP-1 and RNP-2 consensus sequences.

Other human proteins, apart from snRNP U1A, listed in section A which have well conserved RNP-1 and RNP-2 consensus sequences are the heterogeneous nuclear ribonucleoproteins A1 and C1. A1 has two domains (Cobianchi *et al.* 1986 and Kumar *et al.* 1986) and C1 contains one domain (Swanson *et al.* 1987). A summary about the hnRNPs is given below.

The *Drosophila* RNA-binding proteins listed in section A are Hrb98DE and Hrb-87F which are homologous to the human hnRNP proteins with Hrb98DE (Haynes *et al.* 1987 and 1990) being homologous to human hnRNP A1 protein and Hrb87F related to the human A and B hnRNP protein group (Haynes *et al.* 1991).

The proteins Sxl1, Sxl2 (Sex-lethal) and tra-2 (transformer-2) are *Drosophila* proteins with an important role in the sex-determination pathway hierarchy, which is regulated by alternative pre-mRNA processing. The RNA-binding domains of Sex-lethal (Bell *et al.* 1988 and Inoue *et al.* 1990) regulate the choice between two alternative 3' splice sites in transformer (tra) pre-mRNA by binding to a U-rich region upstream of exon 2 and therefore preventing it from being spliced. Tra-2 (Amrein *et al.* 1988 and Goraslki *et al.* 1989) on the other hand is a positive regulator of sex-specific splicing and polyadenylation (Hedley and Maniatis 1991) and regulates the expression of *dsx* (*doublesex*). The binding site for tra-2 in this case is a region containing six copies of a 13 nucleotide repeat found within the female specific exon of *dsx*.

The Drosophila elav protein (elav=embryonic lethal, abnormal visual system; Robinow *et al.* 1988) functions in the maturations of neurons for the development of the visual system. It contains three RNA-binding domains whose binding sites are unknown.

Figure 23b: Alignment of RNA-binding domains

A

```
               RNP-2

               ------

             11* *           *   *   *     *      48
$ snRNPU1A 1  TIYINNLNEKIKKDELKKSLYAIFSQFGQILDILVSRS...............

$ hnRNP A1 1  KLFIGGLSFE.TTDESLRSHFEQWGTLTDCVVMRDPNT...............
$ hnRNP A1 2  KIFVGGIKED.TEEHHLRDYFEQYGKIEVIEIMTDRGS...............
$ hnRNP C1    RVFIGNLNTLVVKKSDVEAIFSKYGKI..VG.CSVH.................

# Hrb 98   1  LFIGGL.DYR.TTDENLKAHFEKWGNIVDVVVMKDPRT...............
# Hrb 98   2  LFVGAL.KDD.HDEQSIRDYFQHFGNIVDINIVIDKET...............
# Hrb 87   1  LFIGGL.DYR.TTDDGLKAHFEKWGNIVDVVVMKDPKT...............
# Hrb 87   2  LFVGGL.RDD.HDEECLREYFKDFGQIVSVNIVSDKDT...............

# Sxl 1       NLIVNYLPQD.MTDRELYALFRAIGPINTCRIMRDYKT...............
# Sxl 2       NLYVTNLPRT.ITDDQLDTIFGKYGSIVQKNILRDKLT...............
# Tra-2       GVFGLNTN...TSQHKVRELFNKYGPIERIQMVIDAQT...............

# Elav1    1  NLIVNYLPQT.MTEDEIRSLFSSVGEIESVKLIRDKSQVYIDPLNPQAPSKGQ
# Elav2    2  NLYVSGLPKT.MTQQELEAIFAPFGAIITSRILQNAGND..............
# Elav3    3  PIFIYNLAPE.TEEAALWQLFGPFGAVQSVKIVKDPTT...............
              |____|       |__________|    |____|
              Beta-1         Alpha-1       Beta-2
```

B

```
               * *            *  *   *     *
$ PBT-4    1  VIHIRKLPID.VTEGEVISLGLPFGKVTNLLMLKG..................
$ PBT-4    2  RIIVENLFYP.VTLDVLHQIFSKFGTVLKIITFTKN.................
$ PBT-4    3  VLLVSNLNPERVTP...QSLFILFGVYGDVQRVKIL.................
$ PBT-4    4  KLHLSNIPPSVSEEDLKVLFSSNGGVVKGFKFFQKD.................

$ hnRNP L  1  VVHIRGLIDG.VVEADLVEALQEFGPI.SYV.VVMPK................
$ hnRNP L  2  LFTILNPIYS.ITTDVLYTICNPCGPVQRI..VIFRK................
$ hnRNP L  3  VLMVYGLDQSKMNGDRVFNVFCLYGNVEKVK.FMKSK................
$ hnRNP L  4  VLHFFNAPLE.VTEENFFEICDELGVKRPSSVKVFSGK...............

# smooth   1  LFTIINPFYP.ITVDVLHKICHPHGQVLRI..VIFKK................
# smooth   2  VMMVYGLDHDTSNTDKLFNLVCLYGNVARIK.FLKTK................
# smooth   3  ILHFFNTPPG.LTEDQLIGIFNIK.DVPATSVRLFPLK...............
```

A

```
                  RNP-1
                  --------

              49   * * * *     *  **  *           * * *    91
$ snRNPU1A 1  .LKMRGQAFVIFKEVSSATNALRSMQGFPFYDKP..MRIQYAKTDS

$ hnRNP A1 1  .KRSRGFGFVTYATVEEVDAAMNARPHKVDGRVV.EPKRAVSREDS
$ hnRNP A1 2  .GKKRGFAFVTFDDHDSVDKIVIQKYHTVNGHNC.EVRKALSKQEM
$ hnRNP C1    ....KGFAFVQYVNERNARAAVAGEDGRMIAGQV..LDINLAAEPK

# Hrb 98   1  .KRSRGFGFITYSHSSMIDEAQKS.RPHKIDGRVVEPKRAVPRQDI
# Hrb 98   2  .GKKRGFAFVEFDDYDPVDKVVLQKQHQLNGKMVDVKKALPKQN..
# Hrb 87   1  .KRSRGFGFITYSQSYMIDNAQNA.RPHKIDGRTVEPKRAVPRQEI
# Hrb 87   2  .GKKRGFAFIEFDDYDPVDKIILQ.KTHSIKNKTLDVKKAIAKQDM

# Sxl 1       .GYSFGYAFVDFTSEMDSQRAIKVLNGITVRNKR..LKVSYARPGG
# Sxl 2       .GRPRGVAFVRYNKREEAQEAISALNNVIPEGGSQPLSVRLAEEHG
# Tra-2       .QRSRGFCFIYFEKLSDARAAKDSCSGIEVDGRRIRVDFSITQR..

# Elav1    1  .GQSLGYGFVNYVRPQDAEQAVNVLNGLRLQNKT..IKVSFARPSS
# Elav2    2  .TQTKGVGFIRFDKREEATRAIIALNGTTPSSCTDPIVVKFSNTPG
# Elav3    3  .NQCKGYGFVSMTNYDEAAMAIRALNGYTMGNRV..LQVSFKTNKA
                   |_____|  |_________|           |____|
                   Beta-3     Alpha-2             Beta-4
```

B

```
                   * * * *     *  **  *           * * *
$ PBT-4    1  ....KNQAFIEMNTEEAANTMVNYYTSVTPVLRGQPIYIQFSNHKE
$ PBT-4    2  ...NQFQALLQYADPVSAQHAKLSLDGQNIYNACCTLRIDFSKLTS
$ PBT-4    3  .FNKKENALVQMADGNQAQLAMSHLNGHKLHGKP..IRITLSKHQN
$ PBT-4    4  ....RKMALIQMGSVEEAVQALIDLHNHDLGENHH.LRVSFSKSFS

$ hnRNP L 1   ....KNQALVEFEDVLGACNAVNYAADNQIYIAGHPAFVNYSTSQK
$ hnRNP L 2   ...NGVQAMVEFDSVQSAQRAKASLNGADIYSGCCTLKIEYAKPTR
$ hnRNP L 3   ....PGAAMVEMADGYAVDRAITHLNNNFMFGQKLNVCVSKQPA.A
$ hnRNP L 4   .SERSSSGLLEWESKSDALETLGFLNHYQMKNPNGPYPYTLKLCFS

# smooth  1   ...NGVQAMVEFDNLDAATRARENLNGADIYAGCCTLKIDYAKPEK
# smooth  2   ....EGTAMVQMGDAVAVERCVQHLNNIPVGTGGK.IQIAFSKQNF
# smooth  3   .TERSSSGLIEFSNISQAVLAIMKCNHLPIEGKGTKFPFIMKLCFS
```

Section B contains alignments of proteins showing only a limited homology to the RNP consensus sequences, but which nevertheless are involved in RNA-binding and have been isolated as nucleic acid binding proteins, binding RNA and single-stranded DNA. For the PTB-4/hnRNP I and the hnRNP L protein, four sequence segments have been aligned and for the smooth protein three.

Not all of these segments are necessarily involved in RNA-binding, as is the case for snRNP U1 A protein, where only one of the two RBDs has a function in RNA-binding. On the other hand, the mammalian splicing factor U2AF (U2 auxiliary factor) contains three RBDs, all of which exhibit a strong RNA binding affinity and play an important role in the RNA-binding (Zamore *et al.* 1992).

As noted before for other RNA-binding proteins with more than one putative RNA-binding domain, domain 1 of PBT-4/hnRNP I is more homologous to domain 1 of hnRNP L than to its other three domains. The same applies for hnRNP L and also for the smooth protein, where the first domain is more homologous to domains 2 of PBT-4 and hnRNP L than to domains 2 and 3 of the smooth protein.

5.2.2 Features and structure of RNA-binding domain

Due to crystallography and NMR studies carried out on snRNP U1 A protein and with more information recently emerging on the global folding pattern of hnRNP C1 also examined by NMR (Wittekind *et al.* 1992 and Görlach *et al.* 1992), amino acid residues which are important for structure have been identified, as opposed to other residues important for function.

As indicated in figure 23b, the secondary structure has been identified, showing a $\beta\alpha\beta - \beta\alpha\beta$ pattern with four antiparallel β-strands and two α helices, whereby the RNP 1 and RNP 2 lie side by side in the middle of two β-sheet strands (β 1 + β 2) (Nagai *et al.* 1990).

It had been observed that the RBDs are rich in hydrophobic and aromatic residues. Certain hydrophobic residues have now been assigned to contribute to the structure of the protein, forming a hydrophobic core (marked by asterisks in fig. 23b), whereas some of the aromatic residues have a function in the RNA-binding.

According to Nagai *et al.* (1990) and Hoffman *et al.* (1991), the hydrophobic core of the RBD domain of snRNP U1 A protein consists of Ile 12, Ile 14, Ile 40, Ala 55, Val 57, Phe 59, Met 82 on the β strands, Leu 26, Leu 30, Phe 34 on α-1 and Ala 65, Ala 68 and Met 70 on α-2.

These hydrophobic amino acids are not only preserved at the above mentioned positions in snRNP U1 A and in section A, but also in section B, allowing hydrophobic amino acids to be replaced by other hydrophobic amino acids.

In the following, the hydrophobic positions of snRNP U1 A and section A in general are compared with amino acid residues in the corresponding positions of section B.

The two Ile 12 and 14 of snRNP U1 A are all occupied by hydrophobic residues at the equivalent positions of section B. Although four Phe residues are present in section B at these positions two each in hnRNP L and smooth.

The next marked hydrophobic position 26, a Leu, in snRNP U1 A does correspond to position 27 according to the U1 A alignment for all the other proteins of section A and B. The Leu is prevalent in the majority of cases, but is occasionally replaced by Ile and Val, with the exception of a Lys in PBT-4.

At position 30, with the Leu in U1 A, a range of hydrophobic residues can be found in section B, including three Leu, five Ile, one Val, one Ala and one Phe. In section A, three His can be found at that position, one of which is in hnRNP A1 1 and the other two in Hrb98 1 and Hrb87 1.

Again position 34 is aromatic with two exceptions, which are still hydrophobic, in section A and five exceptions in B, two of which are not hydrophobic (a His and a Lys).

The corresponding position to amino acid residue 40 again shows a variety of hydrophobic amino acids, whereby section A has one Lys and one Ser as a replacement, there is also a Ser and a Thr to be found in section B. Two aromatic hydrophobic residues are present, one in PBT-4 and the other in hnRNP L 1.

Position 65 contains an Ala in most cases, which holds for section A and B. If no Ala is present, it is either replaced by a Val (in two sequences of section B and four in section A) or by an Ile (two cases in A). Only *Sxl1* has a Ser in that position. Again, position 68 is very well preserved as an Ala with only two replacements by a Val and Ile for section A and three replacements by a Met, a Thr and a Cys in section B. In snRNP U1A, Ala 68 has been shown to be completely buried by the hydrophobic core and is close to two aromatic rings.

The hydrophobic position next to 68 shows mainly hydrophobic residues, with two Gln and one Lys in section A and two Lys and one Arg in section B.

The last two hydrophobic core positions identified in Hoffman *et al.* (1991) and located on β strand 4, are the Met 82 and the Ile 84. Section B has only hydrophobic residues in postion 82, two of which are aromatic residues, whereas three prolines are present in that position in section A. Again, position 84 is very well preserved in section B, with the same two aromatic residues present in the corresponding sequences and section B shows three Arg and two Lys with the remaining residues being hydrophobic, one of which aromatic.

The above has shown that the hydrophobic residues, which were found to have structural importance, are very much conserved even amongst the "nonconsensus" RNA-binding domains.

5.2.3 Comparison of smooth and hnRNP L proteins with regard to the binding domains

The smooth protein and the human hnRNP L protein have got several features in common. The amino termini of both proteins are glycine rich. In the hnRNP L protein, almost 50% of the first 60 amino acids are glycines (60 out of 29). The more than average number of glutamines present at the amino terminus of the smooth protein is not found in the hnRNP L protein.

Another amino acid which is also present at a high frequency almost throughout both proteins is proline, which might be of structural importance conferring flexibility to the proteins.

With a protein sequence of 558 amino acids, a molecular weight of 60.2 and a neutral isoelectric point, the human hnRNP L protein is larger than the smooth protein which has a peptide sequence of 475 amino acids, a molecular weight of 51.9 and an isoelectric point of 8.74.

Sequence comparisons over the whole length of the proteins (see Appendix) reveal three highly conserved regions. These three regions of striking homology are thought to exhibit some functional properties, a possible involvement in the RNA-binding, explaining such great degree of conservation between the invertebrate and the vertebrate protein.

All three regions of the protein alignments listed below (see figure 24) contain either partial or almost complete homologies to the RNA recognition motifs. As suggested by Kenan *et al.* (1991), the hnRNP L protein may contain at least three regions homologous to this motif (if not even four, Ghetti *et al.* 1992). In figure 23b, the four RRM repeats for the hnRNP L protein were listed.

When the first RNA-binding domain of the hnRNP L protein (63-147) is compared to different sections of the *smooth* protein, a significant amount of similarity can be obtained between amino acids 63-147 of hnRNP L and each of the three regions of putative RNA-binding domains of *smooth*. Nevertheless, a direct comparison of the two whole length peptide sequences using the GAP program aligns the second putative RNA-binding domain with the first RBD of the *smooth* protein (see Appendix).

The first and the longest region of extensive overlap between the two peptide sequences covering 93 amino acids (including one gap in the hnRNP L protein), also shows the highest percentage of identity and similarity.

As proposed by Kenan *et al.* (1991), the second RNA-binding domain of hnRNP L would start at amino acid 155 and continue up to amino acid 241. The equivalent regions for the smooth protein would contain amino acids 74 to 178.

The region of homology does not include the first ten amino acids of each protein, which according to the snRNP U1A protein alignment (Kenan *et al.* 1991), would be part of the RNA-binding domain. At the other end, the regions of overlap extend for a further 14 amino acid residues (for the hnRNP L protein) and 15 for the smooth protein.

The second region of homology covers a length of 75 amino acids. Again, this region is part of what could be considered a RNA-binding protein. Both proteins are being matched starting at the equivalent position of amino acid ten of the snRNP U1A RNA-binding domain.

The third region of homology, comprising 69 amino acids, shows least homology to the RNA-binding domian. A putative hnRNP L RNA-binding domain would extend from approximately position 461 to 552 and the aligned *smooth* protein from 353 to 443. The actual region of homology starts 12 amino acids before that and ends 35 amino acids before the end of the putative RNA-binding domain.

As already mentioned in the previous section, the hnRNP L protein and also the smooth protein only contain very loosely conserved RNP consensus sequences (if at all). The amino acids which are specifically conserved in the RNA-binding domains of other RNA-binding proteins, RNP 1 and RNP 2, consisting mainly of aromatic amino acids and amino acids with basic side chains, are absent from both the hnRNP L and the smooth protein (for example, a Tyr or Phe at residue 13 of RNP 2, in RNP 1 a well conserved Arg at position 52, a Tyr or Phe at 54 and a Phe at 56 (all numberings are based on the snRNP U1A RNA-binding domain, Kenan *et al.* 1991 and also used in figure 23b)) .

Therefore the "RNA-binding domains" of the smooth protein and the hnRNP L protein are rather unusual. But this does not necessarily impair the function of either of the proteins as RNA-binding proteins, (which still has to be shown). Kenan *et al.* (1991) argued that, since the major hydrophobic amino acids are maintained in the hnRNP L protein, the folding of the protein would still be characteristic for an RNA-binding protein, where the hydrophobic amino acids make up the core of the RNA-binding domain. A good example of RNA-binding proteins without conserved RNP consensus sequences are the polypyrimidine tract binding proteins (PTB/hnRNP I proteins). Since the PTB proteins were originally purified on their property to bind to RNA, they clearly must be RNA-binding proteins.

Ghetti *et al.* (1992) suggested that, due to their unusual RNA binding domain, hnRNP I/PTB and hnRNP L proteins could form a subfamily of RNA-binding proteins. Although these proteins seem to maintain their 3-dimensional structure of the RRM, the RNA-binding might proceed somehow differently, either with their altered RNA-binding domain or with other regions of the protein outside the RRM.

The other characteristic of this subset of proteins (i.e. the hnRNP I/PTB and the hnRNP L protein) is their antibody staining pattern, which is different to that observed with the other hnRNP proteins.

One final point of interest worth mentioning with regard to the putative RNA-binding domains in the smooth and the hnRNP L protein is that each RNA-binding domain of the smooth protein is more closely related at the level of protein homology to its hnRNP L protein RNA-binding domain counterpart than the other two putative RRMs of smooth itself. The same applies for the hnRNP L protein (see Appendix for the comparisons). A similar situation has been reported for other human and *Drosophila* proteins which contain more than one RRM.

Figure 24: Three regions of high homology between the hnRNP L protein and the smooth protein

Regions of homology between hnRNP L protein and the smooth protein

1) hnRNP L protein × smooth protein

Percent Similarity: 83.696 Percent Identity: 71.739

```
                   .         .         .         .         .
    164 LLFTILNPIYSITTDVLYTICNPCGPVQRIVIFRKNGVQAMVEFDSVQSA 213
        |||||:||:|.||.|||..||:| |.| |||||:|||||||||||.::.|
     83 LLFTIINPFYPITVDVLHKICHPHGQVLRIVIFKKNGVQAMVEFDNLDAA 132

                   .         .         .         .
    214 QRAKASLNGADIYSGCCTLKIEYAKPTRLNVFKNDQDT.WDYT 255
         ||:..|||||||.|||||||:||||.:|||:||:.|| ||||
    133 TRARENLNGADIYAGCCTLKIDYAKPEKLNVYKNEPDTSWDYT 175
```

2) hnRNP L protein × smooth protein

Percent Similarity: 72.973 Percent Identity: 52.703

```
                   .         .         .         .         .
    352 VLMVYGLDQSKMNGDRVFNVFCLYGNVEKVKFMKSKPGAAMVEMADGYAV 401
        |:||||||:.. |.|::||:.||||||.::||:|.|.|.|||:|:|: ||
    242 VMMVYGLDHDTSNTDKLFNLVCLYGNVARIKFLKTKEGTAMVQMGDAVAV 291

                   .         .
    402 DRAITHLNN.NFMFGQKLNVCVSKQ 425
        :|.: ||||  . |.|:.:..|||
    292 ERCVQHLNNIPVGTGGKIQIAFSKQ 316
```

3) hnRNP L protein × smooth protein

Percent Similarity: 75.000 Percent Identity: 54.412

```
             .         .         .         .         .
449 SRNNRFSTPEQAAKNRIQHPSNVLHFFNAPLEVTEENFFEICDELGVKRP 498
    |:|||| .|.||.|||||.||.:|||||.| ::||:.:::| :  :| .:
341 SKNNRFLSPAQASKNRIQPPSKILHFFNTPPGLTEDQLIGIFNIKDV.PA 389

             .
499 SSVKVFSGKSERSSSGLLE 517
    .||::|. |.|||||||:|
390 TSVRLFPLKTERSSSGLIE 408
```

Chapter 6

DISCUSSION

6.0.4 Summary of the original experiment

The primary aim of the original experiment by Mackay (1984; 1985) was to test whether transposable P element mutagenesis can cause a response in quantitative traits. The discovery of an allele of major effect led to the isolation of a *Drosophila* gene. This gene was considered of special interest due to the particular way the mutation had been obtained, i.e. during a selection experiment for a quantitative character. Since matters concerning the nature of genes influencing quantitative characters are poorly understood, characterising this gene of major effect was thought to provide some insight into the question of what kind or type of gene might be expected from a so called "quantitative trait gene".

The mutation isolated exhibited an almost complete absence of abdominal bristles and was associated with a deleterious fitness effect. By virtue of its origin in a dysgenic cross it was believed to have been caused by the insertion of a P element.

Initial mapping experiments located the position of the mutation to approximately 2-80. Subsequent consultation of Lindsley and Grell (1968) revealed that at position 91.5 on the second chromosome a mutation called *smooth*, which closely resembled the new mutant phenotype, had previously been discovered by Bridges and Brehme (1944). A complementation test confirmed that a new *smooth* allele had been found.

Cytological mapping of P elements in extracted second chromosome lines, together with complementation tests, were conducted, using the first *smooth* allele (Bridges and Brehme 1944) in order to score the phenotype of each line. This led to the association of the mutation with one particular P element insertion site, at 56E IIR. Therefore it was believed that the P element had disrupted the *smooth* gene on insertion.

In this thesis I have dealt predominantly with the molecular analysis of the *smooth* locus and described the discovery of a previously unknown *Drosophila* protein coding sequence. The *smooth* gene was found to encode a protein homologous to the human heterogeneous nuclear ribonucleoprotein L protein (hnRNP L), an unusual member of a group of RNA-binding proteins.

Evidence for the P element insertion into the *smooth* gene

The isolation of the abdominal bristle allele of major effect, its mapping and the location of the P element was described earlier (see Introduction and chapters 2 and 4). It had been assumed that the new mutant allele had been caused by the insertion of a P element in a site on the second chromosome which had previously been unoccupied by a P element. The possibility of the presence of a second chromosome carrying a P element at 56E in the original P strain (Harwich), at the start of the experiment, can not completely be ruled out. A chromosome carrying a P element insertion at 56E might have been segregating at a low frequency. But by this stage it has become impossible to deduce whether this had actually been the case. 25 to 30 generations after the second chromosome extraction a small number of P elements at 56E IIR existed in the population.

Reversion experiments resulting in the precise excision of the P element at 56E IIR in at least two cases caused the complete reversion to the wild-type phenotype (A. E. Shrimpton unpublished data). The absence of P elements at the position of 56E after the reversion, as determined by cytological *in vitro* hybridisation mapping, supported the hypothesis of having located the precise position of the P element.

The reversion experiment did not generate any new alleles by imprecise P element excision (A. E. Shrimpton personal communication).

In order to confirm the precise position of the P element by genetic mapping, since the initial mapping had been rather crude and had only yielded a map position of 2-80, Shrimpton proceeded to carry out a fine genetic mapping study. He used multiple flanking markers at loci more closely linked to 2-91.5 than previously (see chapter 3). By scoring the recombinants for their *smooth* phenotypes and establishing the corresponding presence or absence of the P element by *in situ* hybridisation, he found that the P element could not be recombined away from *smooth* and concluded with 95% confidence that the P element was indeed located within one map unit of the *smooth* locus.

Since it had been reported in a number of cases that P elements have a tendency to insert in close proximity to the 5' end or upstream of a gene, this possibility was taken advantage of in this study. Northern blots probed with a fragment flanking the P element from a wild-type chromosome, subsequently showed that the P element had indeed inserted into a transcribed region. This was the only transcript detected within the proximity of the P element insertion site, since further Northern blot analysis ruled out the presence of a transcript upstream of the P element insertion, i.e. upstream of the start of the gene as it was established later.

The isolation of a cDNA clone and the subsequent analysis of the organisation of the gene followed. In total, the gene was found to span a region of at least 90 kb and included 10 exons. The P element had inserted into the first intron of a gene. It was found to be located at nucleotide position 544 from the supposed start of the gene, with the first exon being 466 nucleotides long. The direction of transcription for the P element and the *smooth* gene are the same.

Further support for having cloned the actual *smooth* gene came from preliminary genomic Southern blots, where digests of the three mutant alleles were compared to those of the wild-type blots (results not shown). Various probes derived from wild-type genomic DNA from regions equivalent to the P element insertion site and other flanking regions, showed that there were differences in the region of interest. Both the *sm* allele and *sm*^{2} appeared to be different around the area of P element insertion, showing some sort of rearrangements or insertions. Precise restriction maps of those two alleles were not established.

Although there is strong evidence indicating that the *smooth* gene has been cloned, the fact that the P element is present in the first intron of the gene, causing a mutant phenotype, is rather unusual.

A P element reversion experiment led to the recovery of the wild-type phenotype. From this follows that the inserted P element must have caused the mutant phenotype observed. Since the P element is inserted 120 bp downstream of the first exon, the mutation could be brought about by the P element interfering with the transcription of the gene, preventing, for example, the correct splicing.

On the other hand, it has not been shown whether there are any other transcripts present in the region immediately downstream, adjacent to the P element insertion (i.e. in intron 1). A gene overlapping with the putative *smooth* gene could be responsible for the phenotype observed.

Further Northern analysis, using genomic sequences downstream of the P element insertion as probes, would be necessary to clarify this matter.

6.0.5 *smooth*

The *smooth* gene

Molecular analysis of the *smooth* gene revealed a number of special features. The most immediate observation is concerned with the length of the gene. The *smooth* gene has been found to extend over more than 90 kb of genomic DNA. In contrast, the average length of a *Drosophila* gene typically ranges between 3-5 kb. On the other hand, there are several examples of extremely long and structurally extremely complex genes. For example, the *dunce* locus, which codes for the structural gene cAMP phosphodiesterase, spans a region of at least 150 kb (Qiu *et al.* 1991). It includes 13 exons, eight to ten different transcripts between 4.2 and 9.6 kb in length and one intron 79 kb in length. The *Antp* locus is another example of a long gene. It is also a very complex gene with two promoters and nested transcription units with the longest being 103 kb and the longest intron of 57 kb (Laughon *et al.* 1986).

As observed in genes like *dunce* and *Antennapedia*, the introns in the *smooth* gene are also of varying sizes, with one of them of only 60 bp and others larger than 20 kb. Two of the introns have not been mapped, due to the lack of overlapping lambda clones, so the precise length of all introns is not known.

As already pointed out, the *smooth* gene comprises ten exons, ranging in size from 64 to 590 bp. The first exon is completely noncoding and translation only starts in the last 149 bp of the second exon, the longest, which extends over 590 bp.

The 5'-noncoding region or leader sequence is thought to have a length of 907 nucleotides. The precise 5' end of the transcript has not been mapped experimentally, but there is some other evidence that a complete 5' end of the transcript has been isolated. This indirect evidence is based on the observation that a large number of cDNAs isolated from N. Brown's cDNA libraries show an additional G residue at the 5' end. This artifact which is present in the *smooth* cDNA, is thought to arise when the reverse transcriptase is attempting to transcribe the mRNA cap (see chapter 4).

Interesting observations have been made regarding the 5' end of *Drosophila* transcripts in general. While most of the *Drosophila* genes have leader sequences varying between 40 and 80 nucleotides in length, as many as 20% of *Drosophila* genes contain 5'-noncoding regions (NCRs) of several hundred nucleotides, containing upstream AUGs (Cavener and Ray 1991). This phenomenon has been shown to be very rare for vertebrate genes where, according to Kozak (1991), only genes involved in growth control have long leader sequences.

The *smooth* gene has two translation initiation start sites upstream of the presumed start site. The two AUG sites could theoretically code for two short open reading frames of 28 and 11 amino acids.

A ribosome scanning model for the initiation of translation has been proposed by Kozak (1984; 1989; 1991), which suggests that ribosomes move along the transcript proceeding from the 5' end of the gene and initiate transcription as soon as they would come across the first AUG. In the case of the *Antennapedia* (*Antp*) transcript, the initiation of transcription of the longer transcription unit only starts at the sixteenth AUG and the translation start site is located at nucleotide residue 1727 (the shorter transcription unit starts at nucleotide 1512).

The *Ultrabithorax* (*Ubx*) gene (Kornfeld *et al.* 1989) represents another example of a gene with a long 5' untranslated sequence and in some aspects it is structurally similar to the *smooth* gene. The NCR of the *Ubx* gene closely resembles that of the *smooth* gene in a number of features: the 5'NCR of *Ubx* is 965 bp long, the *smooth* 5'NCR 907 bp, both *Ubx* and *smooth* have two non-consensus AUGs contained in their leader sequence and where *Ubx* has 47 stop codons in that region, *smooth* has 38 stops.

Translation of genes like *Antp*, *Ultrabithorax* and *smooth* would be very inefficient, requiring several reinitiations of the ribosome complex. As has been mentioned earlier, long untranslated regions which include a number of start codons and stop codons could act as a kind of translational regulatory mechanism. Recently, some research has been carried out on trying to find out how efficiently translation could be initiated in these transcripts containing long NCR.

Because the scanning process would be extremely inefficient in these genes, an internal ribosome binding mechanism has been proposed (Oh *et al.* 1992), where translation is initiated from inside a long 5'NCR. This mechanism was first discovered in viruses (Pelletier and Sonenberg 1988; Jang *et al.* 1989) and has now been shown, for the first time, also to function in the translation initiation of the *Antp* gene in *Drosophila.* It has also been reported to be the case in the *Ubx* gene. Oh *et al.* (1992) proposed that many regulatory genes with long 5'NCRs might start translation by internal ribosome binding and identified a number of developmentally important *Drosophila* genes with long leader sequences and multiple upstream AUGs.

Although little more is known about the regulation of translation of the *smooth* gene, the parallels with such developmentally regulated genes are very interesting. It should be mentioned that the selection of the *smooth* mutation was on the basis of an adult phenotype and many *Drosophila* mutations with adult cuticular phenotypes map to genes with more pervasive roles, e.g. *hairy* which has multiple functions during development (see also below).

The smooth protein

Translation of the *smooth* gene sequence with the aid of a computer yielded an amino acid sequence with significant homology to a human protein, which has been described as a novel heterogeneous nuclear ribonucleoprotein L protein (hnRNP L). Even though the precise function of the human protein has not been established, its likely role is as an RNA-binding protein involved in some aspect of pre-mRNA processing. The human homologue, the hnRNP L protein, was found to have a unique distribution on nascent transcript, as indicated by staining of giant loops of the amphibian lampbrush chromosome (Piñol-Roma *et al.* 1989).

Direct comparisons of the protein sequences encoded by *smooth* and the hnRNP L led to the identification of three regions of striking homology with 71.7%, 52.7% and 54.4% identity and 83.7%, 73.0% and 75.0% similarity, respectively (5' to 3'). This is a remarkable degree of conservation between the human and the fly protein in those three regions. The percentage identity over the whole extent of the two proteins amounts to 44%, with the similarity score of 64.4% when using the GAP program (see chapter 4). These numbers obtained are similar to the percentage identity found between the human A-B hnRNP group proteins and the *Drosophila* homologues the Hrb98DE and Hrb87F. Haynes *et al.* (1990; 1991) calculated an identity score of 54% and 59% between the A-B hnRNP proteins to the Hrb98DE and Hrb87F proteins. In view of the degree of divergence between the human A-B hnRNP proteins and the two homologous *Drosophila* proteins, it is likely that the smooth protein is the *Drosophila* the homologue of the human hnRNP L protein.

The homology between the smooth protein and hnRNP L protein extends beyond the three very homologous regions in as far as that both proteins possess a glycine-rich amino-terminus. The main difference persisting between the smooth protein and the hnRNP L protein is that the latter is 83 amino acids longer, corresponding to approximately the length of one of those homologous regions.

Since there is a higher degree of homology in three particular regions, what significance do these regions have and is this high degree of similarity reflecting a functional similarity between the two proteins? At this stage, only speculations can be made, emphasising that the similarities are purely based on sequence homology.

As mentioned earlier, the hnRNP L protein is part of the hnRNP complex and is regarded as one of the major hnRNP proteins, but considered to be a rather exceptional member of the group of hnRNPs. Many of the human hnRNP proteins belong to a family of RNA recognition motif (RRM) containing proteins. This particular motif has been found in a large number of evolutionary divergent species i.e. from vertebrate species to invertebrate species, including *Drosophila*, yeast (Adam *et al.* 1986; Sachs *et al.* 1986), *C. elegans* (Iwasaki *et al.* 1992), brine shrimp (Thomas *et al.* 1983) and grasshopper (Ball *et al.* 1991). A consensus sequence has been established which contains many hydrophobic and aromatic amino acid residues at strategic positions and the two RNP elements. This high degree of amino acid conservation found in many of the domains, suggests an RNA-binding function.

When examining the putative RNA-binding domain (RBD) of the hnRNP L protein, it can be observed that it does not fit the consensus sequence very well; in particular, the lack of the usually highly conserved RNP elements becomes apparent. The hnRNP L protein has been described as being different with regard to the RBDs, because it possesses unusual RNA recognition motifs, although it is still recognisable as such.

Another example of an hnRNP protein containing unusual RRMs is the hnRNP I/PTB (pyrimidine-tract binding) protein. Both the hnRNP L and the hnRNP I/PTB proteins have been shown to be capable of binding single-stranded nucleic acids. While the hnRNP I/PTB protein binds pyrimidine tracts, the hnRNP L protein also seems to have a tendency to preferentially bind C-rich regions of transcripts. Nevertheless, there is a distinct difference in the RNA-binding domain between the hnRNP L/hnRNP I/PTB proteins compared to the consensus RBD in the absence of the otherwise highly conserved RNP-1 and RNP-2 consensus sequences, which predominantly consist of aromatic amino acid residues and are thought to be actively involved in the RNA-binding.

How could RNA-binding be achieved by the hnRNP L and hnRNP I/PTB proteins? Evidence from the crystal structure (Nagai *et al.* 1990) and NMR (Jessen *et al.* 1991; Hoffman *et al.* 1992) spectroscopy of RNA-binding domains has shown that RBDs contain a number of critical hydrophobic and aromatic amino acid residues at critical positions. In particular, the hydrophobic residues are of structural importance for the formation of the core of the domain, whereas the aromatic residues play a role in the RNA-binding.

Alignments of the putative RNA-binding domains of hnRNP L/ hnRNP I/PTB proteins and the smooth protein with the general RBD consensus has shown that these hydrophobic amino acids are notably conserved (chapter 5). It has therefore been suggested that the hnRNP L and hnRNP I/PTB proteins could be capable of binding RNA with their lesser conserved RBDs. Their unusual RRMs might confer different binding specificities compared to those of other RRMs. With experimental evidence already presented for specific RNA-binding of the hnRNP I/PTB and hnRNPL proteins, there is a strong argument for the smooth protein also being capable of binding to RNA in a similar fashion, recognising a specific RNA or class of RNA molecules. Hence the smooth protein should be included as a member of the family of RNA-binding proteins.

There is still a possibility that RNA-binding might be mediated by other sequence motifs present in the hnRNP L and hnRNP I/PTB proteins. It has been observed that a large number of RRM containing proteins have unique sequences outside their RRM, many of them abundant in a particular amino acid. Bandzilius *et al.* (1989) suggested that these sequences flanking the RRM , the so-called auxiliary domains, might assist the RNA-binding by affecting the specificity or affinity of it. These domains could also be involved in protein/protein interactions, since the latter are also required for the formation of the hnRNP complex. In a number of proteins, this domain can be recognised by either a high concentration of one amino acid, e.g. Gly in the carboxy terminus of the human hnRNP A1, a series of prolines in the yeast PABP (Adam *et al.* 1986; Sachs *et al.* 1986) or glutamic and aspartic acid residues which can be found in hnRNP C (Swanson *et al.* 1987).

Another group of RNA-binding proteins exist where binding takes place via a series of alternating serine-arginine (S-R) dipeptide residues, predominantly found at the carboxy-terminus of proteins. The so-called SR proteins are often found to possess this SR domain in addition to a RNA-binding domain at the amino-terminus and belong to a conserved family of pre-mRNA splicing factors. But in the hnRNP L and hnRNP I/PTB proteins, a SR dipeptide motif is not present.

Both the hnRNP L and hnRNP I/PTB proteins contain four sequence repeats i.e. the putative RNA-binding domains (only three in the smooth protein). Whereas in the hnRNP I /PTB protein there are no specific features found outside the four RRMs, in both the smooth protein and the hnRNP L protein the amino-termini are glycine-rich. In the smooth protein, the number of glutamines at the amino terminus is also higher than throughout the rest of the protein.

It has been suggested that the amino acid glycine can be involved in protein-nucleic acid or/and protein-protein interactions. *In vitro* experiments carried out with the glycine-rich carboxy terminus of the human hnRNP A1 protein, have shown that this carboxyl end is capable of binding either proteins or nucleic acids (Cobianchi *et al.* 1988; Kumar *et al.* 1990; Nadler *et al.* 1991).

What implications this could have for the amino terminus of the hnRNP L and/or the smooth protein, is a question that still needs to be examined and could be subject to biochemical investigation.

6.0.6 hnRNPs in *Drosophila*

Most of the initial work conducted on hnRNP proteins was carried on human HeLa cells. Very little was known regarding hnRNPs in *Drosophila* and only in recent years the analysis of hnRNP complexes and its individual components has intensified.

A common feature to many hnRNPs is a glycine rich carboxy-terminus (as opposed to the glycine-rich amino-terminus found in the hnRNP L protein and the smooth protein). When Haynes *et al.* (1987) were searching for homologues to glycine-rich *Drosophila* sequences, they isolated a protein, p9, which appeared to have an interesting structure. They found three sequence repeats, each of them displaying sequence similarities to the RNA recognition motifs of characteristic RNA-binding proteins. The protein also contained a glycine-rich carboxy-terminal end.

Further analysis of the sequence indicated that p9 was a homologue of the rat helix destabilising protein (Cobianchi *et al.* 1986) and also of the human hnRNP A1 protein. p9 was subsequently renamed to Hrb98DE (hnRNA binding protein) according to its cytological location, i.e. 98DE. Four isoforms of the protein were isolated (Haynes *et al.* 1990). Subsequent cross-hybridisation experiments using parts of the Hrb98DE cDNA as a probe, led to the isolation of a further member of the A-B hnRNP protein group, Hrb87F (where 87F again refers to the cytological location on the *Drosophila* polytene chromosome). Hrb87F contains two RNA-binding domains and a glycine-rich COOH-terminal region. Hrb98DE and Hrb87F are 76% identical over 557 nucleotides in their RNA recognition motifs.

Further investigation into the group of hnRNP proteins in *Drosophila* was carried out by studying the hnRNP complexes as a whole, attempting to compare the hnRNP complexes of *Drosophila* to those of humans. This work was stimulated by the fact, that among vertebrates, a very high degree of conservation was found between different species (>90%), e.g. *Xenopus* and human hnRNP A1 (Haynes *et al.* 1991). Is this degree of conservation maintained between vertebrates and invertebrates?

When the protein sequences of *Drosophila* hnRNP Hrb98DE and Hrb87F proteins were compared to those of their human counterparts, the A-B hnRNP proteins, an identity score of 54 and 59% was obtained for the Hrb98DE and Hrb87F, respectively (Haynes *et al.* 1991). This degree of homology is less than had been anticipated, taking into account their supposed similarity of function.

Two research groups, Beyer and colleagues and Dreyfuss and colleagues, have been working on the isolation of the *Drosophila* hnRNP complexes and comparing them to the human hnRNP complexes.

Raychaudhuri *et al.* (1992) prepared antibodies to the RBD domains of Hrb98DE and reacted these to Western blots of nuclear extracts from *Drosophila* Schneider cells. Subsequent 2-D analysis of these immunoblots revealed 10 to 15 distinct polypeptides in the basic region of the gel. The antibodies recognised hnRNPs that were similar in size, charge and number to the group of the human A-B hnRNP proteins, but vertebrate monoclonal antibodies to A-B hnRNP proteins did not react to any proteins of the *Drosophila* hnRNP complex. Hence, it was concluded that there is a substantial degree of divergence between the mammalian and the *Drosophila* hnRNP proteins.

Two alternative hypotheses were proposed by Raychaudhuri *et al.* (1992) regarding these differences. Either invertebrate and vertebrate hnRNPs had evolved from a common ancestor and only the functionally important sequences had remained, while the others had diverged; or convergent evolution could have taken place, resulting in the development of similar RNA-binding regions, but maybe involving varying mechanisms of RNA-binding.

Matunis *et al.* (1992a; 1992b) also succeeded in the isolation of hnRNP complexes from *Drosophila.* Taking advantage of the single-stranded nucleic acid binding property of hnRNPs, proteins were purified from *Drosophila* embryos by single-stranded DNA chromatography. Monoclonal antibodies to these purified proteins were raised and some promising candidates underwent further analysis, using immunofluorescence microscopy to confirm the position of the proteins, expecting them to be located in the nucleus. Polytene chromosome immunofluorescence was also carried out to investigate whether proteins were binding to actively transcribing (puffing) loci on the chromosomes. This led to the identification of a number of authentic hnRNPs. Two of these proteins isolated, pHRP38.1 and pHRP40.1 (and two further isoforms of pHRP40.1: pHRP40.2 and 36.1) were identical to Hrb98DE and Hrb87F, as previously described by Haynes *et al.* (1987; 1990; 1991). A protein called HRP34 was found which appeared to be homologous to the human hnRNP C protein.

Monoclonal antibodies were prepared to HRP40 (heterogeneous ribonucleoprotein) and were used to immunopurify hnRNP complexes. The observations made were similar to those described by Raychaudhuri *et al.* (1992): the hnRNPs in *Drosophila*, as in humans, showed a similar, large and diverse complex of proteins with an abundant number of isoforms. Due to different electrophoretic mobilities, a direct comparison of *Drosophila melanogaster* hnRNP complexes and human hnRNP complexes was not possible.

Matunis *et al.* (1992a; 1992b) managed to isolate *Drosophila* antibodies which did cross-react to hnRNP proteins of HeLa cells and other vertebrate species, e.g. *Xenopus*, chicken and *Saccharomyces cerevisiae* (so far no *S. cerevisiae* hnRNP protein has been described at the molecular level).

On the whole, *Drosophila* hnRNPs were found to be abundant and their composition similar to human hnRNPs. Differences were established in their binding activities according to binding studies. On the other hand, there is likely to be a major requirement for hnRNPs, considering the large number of processes taking place in the cell as part of the RNA metabolism. Whether each isoform of a protein indeed has its own specificity and/or affinity and works in a different fashion to other isoforms, has not been investigated. There might be a high degree of redundancy, considering the number of isoforms which have been shown to be present.

With regard to the conservation of hnRNPs among different species, it is surprising that no hnRNP has been identified in yeast so far. There is a requirement for RNA processing in yeast and splicing factors like the small nuclear RNAs (U1, U2, U4, U5 and U6) have been well characterised. But it has been pointed out (Guthrie and Patterson 1988) that compared to the mammalian snRNAs, the yeast snRNA counterparts are not very well conserved. It has been proposed that this could reflect the differences observed in the splicing apparatus in yeast and mammals, i.e. most mammalian introns are not recognised by yeast (Beggs *et al.* 1980), partly due to differences in the splicing signals. The requirement for other RNA processing factors in yeast might therefore differ significantly, too.

hnRNP and other proteins without RNA recognition motifs

Although, in *Drosophila*, all of the hnRNP proteins described so far contain two RNA-binding domains and a glycine-rich carboxy-domain, among the human hnRNP complex major proteins, a number have been identified which are only distantly related, or not related at all, to the RBD containing hnRNPs. These are the hnRNP I, J, K, L, M and U proteins ((Patton *et al.* 1991; Gil *et al.* 1991; Ghetti *et al.* 1992); K (Matunis *et al.* 1992c); L (Piñol-Roma *et al.* 1989); U (Kiledjian and Dreyfuss 1992)). The only two of these proteins which have a significant similarity to each other are the hnRNP I/PTB and hnRNP L proteins. Both of these exhibit some characteristic features of the RRM and have been shown to bind to oligo(dC) in a highly salt resistant manner (Matunis *et al.* 1992c).

The hnRNP K and hnRNP U proteins were found to be substantially different from the other hnRNPs, showing no evidence of an RRM and each having their own unique features. HnRNP K (Matunis *et al.* 1992c) has been identified as an abundant major poly(C) binding protein and a special feature in its primary structure has been revealed. These are two sets of internal repeats, rich in glycine. Suggestions have been made that these sequence repeats may be involved in RNA-binding.

In the hnRNP U protein (Kiledjian and Dreyfuss 1992), which represents the largest of the hnRNP proteins (120 kD) and an abundant nucleoplasmic phosphoprotein, it has actually been demonstrated that a 26 amino acid region in the glycine-rich carboxy-terminus is capable of binding RNA. This region contains a number of RGG (ArgGlyGly) motif boxes, which have been found in other proteins, e.g. the hnRNP A1 protein.

Since only a small number of proteins of the *Drosophila* hnRNP complex have been characterised which so far have exhibited only similarity to the RBD containing hnRNPs, there is still some scope to find (if they exist) some of the above mentioned types of protein with either atypical RRMs or no RRMs.

Proteins purified by single-stranded DNA chromatography in *Drosophila* (Matunis *et al.* 1992a; 1992b) were subsequently separated by 2-D electrophoresis and revealed most of the major proteins of the hnRNP complex in. Matunis *et al.* (1992a; 1992b) examined six of them in more detail. That still leaves a number of those polypeptides uncharacterised.

The counterpart of the human hnRNP L proteins which could represent the *Drosophila* smooth protein might be amongst them and a protein HRP54, i.e. with a molecular mass of 54 kD, found in the basic region of the gel, might be a good candidate for the gene. The calculated isoelectric point of the smooth protein is 8.74.

Further observations and suggestions

When the human hnRNP L protein was first isolated (Piñol-Roma *et al.* 1989), a Western blot with samples of total cellular proteins of several different species was reacted to anti-L-antibodies. This immunoblot did not detect any signal from the *Drosophila* proteins, compared to the presence of the clear L-protein bands in the tracks of vertebrates like mouse and newts. *Xenopus* was also reported to produce a cross-reaction but no signal was obtained for *Saccharomyces cerevisiae.*

The most probable reason for not detecting any cross-reactivity to any *Drosophila* proteins using the anti-L-antibody was the lack of sensitivity in the test.

Another possible explanation for the absence of any cross-reactivity between smooth and the anti-L-antibody would be that *smooth* is more closely related to the human hnRNP M protein than to the L protein. The hnRNP M protein (Piñol-Roma *et al.* 1988) has approximately the same size as the human L protein, but is slightly more basic, similar to the smooth protein. Unfortunately, this seems to be all the published information available on the human hnRNP M protein at the moment.

The smooth protein could also represent a homologue of the human hnRNP I/PTB protein, which is smaller in size and also more basic than the hnRNP L and hnRNP M protein. But, has been mentioned before, it is not possible to compare the mobilities of the human proteins to those of the *Drosophila* proteins. Additionally, the overall similarity between the mammalian and *Drosophila* hnRNPs has shown to be quite poor. Therefore it is probably not possible to carry out any direct comparisons regarding the human and *Drosophila* hnRNP protein.

It is obvious that further experiments are needed to clarify the matter. It remains to be confirmed that the smooth protein indeed belongs to the group of hnRNP proteins. General methods of confirming the properties of an RNA-binding protein can be applied, like those suggested by Mortensen and Dreyfuss (1989) and include the testing "by chromatography on small-subunit DNA columns, ribohomopolymer columns, and by photochemical cross-linking to RNA".

Immunopurification would be another option, since antibodies could easily be raised to the protein which can be obtained by expressing it in *E. coli.* Immunofluorescence could also be used for checking the localisation of the protein in the cell and investigating polytene chromosomes for actively puffing loci and co-expression of RNA polymerase II at those positions.

Other proteins in *Drosophila* with RRMs

Approximately 25 *Drosophila* genes have been identified so far containing RRMs. Three quarters of these have housekeeping functions e.g. the hnRNPs and snRNPs, whereas the remainder play important developmental regulatory roles, being involved in the splicing of transcripts in a sex-or tissue-specific manner.

The RRM containing proteins which are involved in the sex-determination pathway in *Drosophila*, i.e. Sxl and tra-2, have been shown to carry out alternate splicing processes and can therefore be described as splicing factors.

Another RRM containing protein, the elav (embryonic lethal, abnormal visual system) protein has been characterised (Robinow *et al.* 1988). This protein is involved in the RNA metabolism of the nervous system, probably in the control of tissue, i.e. neuron-specific RNA splicing. A very recently isolated RRM containing *Drosophila* gene, *rbp9*, which is highly similar to *elav*, is also nervous-system specific and probably involved in adult neurogenesis. *rbp9* contains three RRMs (Kim and Baker 1993b). There is also the *suppressor-of-white-apricot* (*su*(w^a)) (Chou *et al.* 1987), which autoregulates its own splicing (Zachar *et al.* 1987).

An increasing number of proteins containing the characteristic RRM are now being isolated in *Drosophila* and most of them are thought to have some sort of developmental regulatory function, but their binding substrates or precise functions are mostly unknown. Other examples of recently isolated RRM containing *Drosophila* non-hnRNP proteins are suppressor-of-sable (su(s)) (Voelker *et al.* 1991), couch potato (cpo) (Bellen *et al.* 1992) and oo18RNA-binding (orb) (Lantz *et al.* 1992). All of these proteins contain at least one RRM, with a more or less conserved RNA-binding domain and all of them exhibit at least a conserved arrangement of hydrophobic residues.

In total, it has been estimated (Kim and Baker 1993a), that there could be approximately 300 RRM-containing proteins in *Drosophila* involved in one way or other in RNA metabolism.

6.0.7 Comparison of the gene structure and the developmental expression of the *Drosophila* hnRNP proteins Hrb98DE and Hrb87F to smooth

The Hrb98DE and the Hrb87F proteins are very similar to each other (79% identity over their complete length according to a GAP comparison) and are thought to be homologous to the human A-B type hnRNP proteins. Both of them show very similar special features.

The Hrb98DE gene is about 6 kb long and consists of eight exons, where in any one transcript there is an alternative first exon, either 1a or 1b and an alternative last exon, either 6a or 6b. Eight different, alternatively spliced transcripts of approximately 2.3 kb in size are present and four different protein isoforms have been detected. The molecular mass of the protein varies between 30-35 kD.

The Hrb87F gene is smaller and only extends over a region of 3.2 kb. Four exons are present, where there is a choice of two polyadenylation sites for the last exon and therefore two different transcripts are produced, 1.7 and 2.2 kb in size, depending on the position of the polyadenylation site.

It was observed for both genes that their splicing pattern is at many positions similar to that of the hnRNP A1 gene, with the same exon/intron splice sites located at equivalent positions of the genes. Also, in all three (i.e. Hrb98DE, Hrb87F and hnRNP A1) of them, the most 3' exon is noncoding which is rather unusual. The developmental pattern of transcription for the Hrb98DE and Hrb87F genes is also very similar. The level of transcription in the early embryo is quite high, but reduces towards the end of embryogenesis. This implies that the genes are maternally transcribed. An increased level of transcription can be observed over the larval period, reaching a maximum at the pupal stage. The behaviour of the smaller Hrb87F transcript of 1.7 kb follows a slightly different pattern, where the level remains the same throughout development and stays at the level reached in the late embryonic stage. Apart from the adult mRNA, which had not been included in any of the Northern blots shown, expression can be found in all developmental stages.

The *smooth* gene, on the other hand, differs from these two genes in several aspects. As mentioned before, the gene is very long for *Drosophila*, extending over more than 90 kb and has ten exons irregularly distributed over this length. Only a single transcript of 2.6 kb has been detected (which does not rule out that others, differentially spliced transcripts of the same length might not be present) and the translation product has a molecular mass of approximately 52kD. It is not possible to compare the splicing pattern or the organisation of the *smooth* gene with its putative human homologue, the hnRNP L protein, since only the cDNA clone has been characterised for the latter. No further information seems to be available at this point in time on the human hnRNP L protein.

The Northern blot of *smooth* reveals different levels of transcripts at different developmental stages. The embryonic, larval and pupal time collection points had not been subdivided as in the case of the Northerns blots for the transcripts of the Hrbs (Haynes *et al.* 1990; Haynes *et al.* 1991), which makes a comparison more difficult. The transcript level of *smooth* in the ovary and 0-12 h embryo has not been determined. There is no detectable level of the 2.6 kb transcript present in the 12-20 h embryo (although a smear can be seen in the embryo lane of figure 17), a low level of transcription in the second instar larvae, a maximum level in the pupae and a probably slightly lower level again in the adult can be observed. Comparing the transcripts of the three proteins (Hrb98DE; Hrb87F and smooth), there is no detectable level of transcription during late embryogenesis in any of them (apart from the smear in *smooth*). An increase at the larval stages can also be observed in all of them. The highest level of transcription can generally be seen in the pupal stage. The adult level of transcript showing a slight reduction from the level observed in the pupae can only be observed in *smooth*, due to the fact that there is no data available for the others.

It can be concluded that, apart from the lower 1.7 kb transcript of Hrb87F, the pattern of transcription seems to be similar in all three genes with regard to the Northern blots available.

Haynes *et al.* (1990) suggested that the Hrb98DE proteins might possibly function as basic housekeeping proteins, but observed that the level of transcription present at each developmental stage was not uniform. If the level of protein expression is correlated to the level of transcription, then varying amounts of proteins are expressed in the cells. On the other hand, the amount of protein present does not need to correspond to the level of transcription and hence it is not possible to deduce from the level of transcription the actual amount of protein expression at different stages during development.

Only the small Hrb87F transcript of 1.7 kb seems to be transcribed at the same level throughout development. It has also be shown that the amount of protein produced by that particular transcript is sufficient for normal development (Haynes *et al.* 1991). The smaller Hrb87F transcript therefore behaves as a classical housekeeping protein.

If the pattern of transcription obtained for the other two Hrb transcripts represents the amount of proteins in the cell, then clearly the level of protein varies immensely between different stages, in particular during late embryogenesis. Alternatively, it could be the case, as suggested by Haynes *et al.* (1991), that enough protein is produced by the mother to suffice for the early development.

Zygotic transcription for the Hrb proteins (and also as suggested by the Northern blot, for the smooth protein) is only switched on at the early larval stage and gradually increases towards the pupal stage. During the larval period, the transcription apparatus must be quite active, since larvae have to undergo a series of moults during that part of development. When the overall level of transcription generally increases, the proteins of the hnRNP complex are very likely to be required in order to mediate the processing of the pre-mRNA.

What could be the cause of the apparent absence of the *smooth* transcript during embryogenesis? It could be speculated that the *smooth* transcript has a more specific function and is only required from the onset of the larval stage. Considering this in the context of abdominal bristle development, which only starts during the pupal stage, the question arises as to why are there any transcripts present during the larval stage. On the other hand, taking into consideration the pleiotropic effect, some of the other bristles also affected by the expression of *smooth* are produced slightly earlier in development than the abdominal bristles which only are the last bristles to develop. But care has to be taken trying to equate the presence of transcription with the presence of gene expression. Therefore, whether the above is a satisfactory explanation for the different level of *smooth* transcripts during development still needs to be confirmed. In particular, with regard to the results obtained for Hrb98DE and Hrb87F (Haynes *et al.* 1990; Haynes *et al.* 1991), the level of transcription of *smooth* might well parallel that of the Hrbs.

There is one more important point, which needs to be considered in the context of the question as to whether there are any transcripts present in the embryos or not.

Due to the fact that the cDNA clone sequenced was derived from a 12-24 h library, the question arises, why has no transcript (of approximately 2.6 kb in length) been detected for the embryo on the Northern blot.

A simple answer could be, that the library is contaminated with larval cDNA. In order to clarify this apparent discrepancy, cDNA libraries of embryos from different early developmental stages need to be probed with the already obtained cDNA clone, checking for the presence or absence of transcripts. Northern blots of egg mRNA from different time points could also be analysed.

Another difference between the Hrb genes and the *smooth* gene is, of course, the fact that mutants are available. The *smooth* mutant, *sm*^{3}, can be described as (semi)-lethal and, probably because it is not a null mutant, a small percentage of embryos develop into viable adult flies. Since the hnRNPs seem to be of functional importance during most stages of development, they are expected to be lethal. In a section below speculations are made, attempting to associate the genotype to the phenotype observed.

6.0.8 What is the relationship between the genotype and the phenotype?

The *smooth* gene has been found to code for a protein (on the basis of sequence identity) which resembles a human hnRNP protein and contains putative RNA-binding domains. A major question, which has not been addressed so far is, how could the protein be related to the observed phenotype, or what is the relationship between the phenotype and the genotype? Further, if the gene isolated is definitely causing the *smooth* mutation, what function can one deduce from the gene product leading to such a mutation?

As described above, since the *smooth* gene is likely to be encoding an RNA-binding protein, the assumption is that it will be involved in some aspect of pre-mRNA/RNA processing which encompasses a large number of functions including splicing, transfer, storing, etc.. To narrow down those functional properties, I have suggested, on the grounds of sequence identity, that the *smooth* gene could represent a homologue to the human hnRNP L protein, which has been shown to bind to actively transcribing regions on the lampbrush chromosome. Unfortunately, the precise function of the hnRNP L protein has not been determined yet, but other hnRNP proteins have been shown to have binding specificities to, for example, the 3' end of the intron.

In the case of the hnRNP I/PTB protein which is more closely related to hnRNP L than any other known hnRNP proteins, binding to the polypyrimidine tract which is closely located to the 3' end of the intron has been observed, hence implying a function in the splicing process.

Before attempting to relate the molecular observations directly to the phenotypic observations, a summarised description of the different phenotypes is carried out below, starting with the original *smooth* mutation (sm^1).

smooth^1 was discovered by Bridges and Brehme (1944). The phenotype is fully described in Lindsley and Grell (1968) (also see chapter 3), but several aspects of the description have to be emphasised and considered in further detail. One important observation is that *smooth*^{1} shows a 30% viability compared to wild-type and the second is that both sexes are entirely sterile. The mutant strain *Cy/sm px* obtained from the stock centre, which was supposedly carrying the original *smooth* mutation, did not prove to be homozygous viable, probably due to the accumulation of lethals. Hence, with regard to the original mutation, the only information available is that provided by Bridges and Brehme (1944).

The *sm*^{2} mutant isolated by Frankham and Nurthen (1981) is completely viable with both sexes being fertile. This *smooth* allele only shows a weak bristle phenotype (see chapter 3).

The *smooth*^{3} allele has been isolated as part of a P element hybrid dysgenesis experiment which required the presence of a large number of P elements on all chromosomes. The line which was initially isolated, *Cy/TR20*, had a number of P element insertions on the chromosomes and therefore the lethal effect observed in this line when homozygous might not exclusively be due to the P element insertion into the *smooth* locus. After outcrossing most of the P elements (see chapter 4) A.E. Shrimpton obtained a line, *Cy/LP2*, which only carried four P element insertions in total, one of which was on the second chromosome i.e. the insertion at the cytological position of 56E in the *smooth* locus. *Cy/LP2* is therefore the most informative line of those lines available containing a P element at 56E IIR, with regard to viability of any of the *smooth*^{3} mutant lines. Crosses between flies of this line result in mostly lethal flies, i.e. hardly any homozygous *LP2/LP2* adult flies emerge. But, due to fact that *Cy/LP2* is still not completely free of other P elements, care has to be taken in concluding that the P element on the second chromosome is entirely responsible for the phenotype observed. With regard to the fertility of the few emerging *LP2/LP2* adult flies, both sexes are sterile.

Comparing the three alleles available and assuming that the P element insertion in *smooth*^{3} is responsible for the observed lethals, the latter is most likely the strongest of the three known *smooth* alleles. The question following this is what can be speculated about the null mutant? Is the *smooth*^{3} allele the null mutant? And if not, what phenotype could be associated with the null mutant.

It is assumed that the *smooth* null mutant would be lethal, due to the fact that the already characterised alleles *smooth*^{1} and *smooth*^{3} are almost lethal. Since apparently the *smooth*^{1} mutant had a much higher percentage of viability compared to *smooth*^{3}, the latter might resemble most closely to the null mutant. But, due to the fact that a small number of viable flies emerge and the abdomen is not completely denuded of its bristles, there is evidence for the *smooth*^{3} mutant being a leaky, rather than a null mutant.

When trying to establish the time of lethality of *smooth*^{3} mutants, taking into account the results of the Northern blots, it can be assumed that sm^3/sm^3 embryos are still viable. Two different lines of evidence can support this assumption: firstly, there is no transcript detectable during embryogenesis (at least between 12–20 hrs of embryogenesis on my Northern blot) and secondly, even if there was a transcript present very early during development, it would most likely be maternally supplied.

The low levels of transcript present at the larval period indicate that there might be a requirement for the smooth protein at the late larval stage or the early pupal stage, since the transcript level increases drastically at the pupal stage (if the transcript level parallels the level of protein expression). From these observations, one can deduce that *smooth* could be a larval or a pupal lethal, hence surviving embryogenesis.

Another point which has to be made is the fact that only the adult *smooth* phenotype is known. But apart from the obvious bristle phenotypes, the homozygous *smooth*^{1} and *smooth*^{3} adult flies have also been shown to be sterile. There could be several causes for male and female sterility, one of which could be associated with defunct mating behaviour. Female sterility is normally due to mutations in genes required for some aspect of oogenesis which, on the other hand, male sterility would result from mutations in spermatogenesis. But, since both sexes are affected, it can speculated that the formation of the pole plasm is somehow prevented, leading to sterility in both sexes. The above therefore indicates that there might also be an early requirement for the *smooth* gene product which could be maternally supplied and would be sufficient until the start of larval development.

This, on the other hand, would imply that *smooth* has more than one function, a not uncommon phenomenon in *Drosophila.* A number of genes have been shown to have multiple developmental functions. One prominent example is the *hairy* gene (Ingham *et al.* 1985), which is required during two distinct developmental stages. Apart from the requirement for bristle development at the adult stage being expressed relatively late in the development, *hairy* is also a pair-rule gene necessary for the segmentation of the embryo (Ingham et al. 1985).

Another well characterised gene with multiple functions is the *engrailed* gene, which is zygotically transcribed with varying roles during development. *engrailed* is involved in segmentation and has been described as a selector gene specifically required to maintain borders between cells (Morata and Lawrence 1975). The original mutant found was a mutation in the adult fly showing a cleft along the middle of the scutellum (hence engrailed) and irregular venation of the wings (Eker 1929).

Concluding from the above, due to the *smooth* phenotype and the molecular observation concerning the long 5' end of the gene showing similarities in many respects to a number of developmental genes, there might be an indication that *smooth* is required early in development, as well as later on.

Possible function of the smooth protein

How can the lack of RNA processing in some way or other cause a mutation in the abdominal bristles of the adult fly, which also reduces viability and causes sterility, apart from other pleiotropic bristle effects? Are there any other examples of genes with known abdominal bristle effects affecting housekeeping genes? Two well-known bristle mutants have been studied for a long time and will be described in more detail below. These two mutants are *bobbed* (*bb*) and *Minute* (*M*). The interesting connection between *Minute* and *bobbed* is that both of them are thought to be structural genes involved in ribosome synthesis.

A mutation in *bobbed* leads to the shortening, thinning and reduction of bristles. It delays the development and causes the etching of abdominal tergites. The *bobbed* gene is present on the X and Y chromosome, and consists of a cluster of about 200 tandemly arranged rRNA genes. It was found to encode the 18S and 28S ribosomal subunits. *bobbed* mutations are often caused by an unequal cross-over event and hence reduce the total number of functional 28S repeats by a large jump. Such an event was described in the course of a selection experiment by Frankham *et al.* (1978). The severity of the phenotype has been shown to depend on the number of functional rRNA genes. And although there are no differences in the RNA contents of oocytes in the mutant compared to the wild-type, the overall effect of the mutation is a slow down in development. One explanation for the bristle effect was delivered by Kay and Jacobs-Lorena (1987), who suggested that due to the delay in the development, during the pupal stage, the "components" of the bristles might not be synthesised in sufficient quantity, hence the thinning, shortening and even reduction in number of bristles.

The *Minute* mutation encompasses a group of haplo-insufficient mutants with more than 50 mutations associated with similar phenotypes, but unlinked and non-additive. The phenotypic characteristics include prolonged larval development, reduced viability, short thin bristles and small cell size, being the cause of the smaller body size compared to normal wild-type size. *Minute* mutants are recessive lethal. An association between the numerous *Minute* mutants and the ribosomal protein genes which are also present in large numbers has been suggested, because they tend to map to the same cytological positions on the polytene chromosomes. Whether this implies that all *Minute* loci code for ribosomal protein genes is not known, but it was postulated to be the case a number of years ago (Kongsuwan *et al.* 1985). As a matter of fact, only one *Minute* locus, at 99D, has been identified so far as being associated with a ribosomal protein, i.e. rp49 (Kongsuwan *et al.* 1985).

Speculations can be made regarding as to whether the *smooth* gene is required to be fully active to allow the proper rapid expression of structural proteins in the bristle cells or ,alternatively, is involved in the series of events leading to the biosynthesis of ribosomal proteins. It could be suggested that *smooth* acts as an RNA-binding protein, being somehow involved in the processing of ribosomal protein transcripts. But the matter gets more and more complex, considering the developmental expression of the ribosomal protein transcripts.

The ribosomal protein transcripts display a specific expression pattern, in particular during the early development. There is an abundance of the maternally supplied ribosomal protein mRNA in early *Drosophila* embryogenesis and it has been reported that 7% of all maternally supplied transcripts are ribosomal.

In spite of those large amounts of ribosomal protein transcripts, the embryo actively translates a number of other maternal transcripts which are present, but the synthesis of ribosomal proteins seems to be discriminated against by some kind of translational regulatory mechanism. During late embryogenesis, ribosomal protein synthesis is activated again (Mager 1987).

What could the role of the smooth protein be?

One possible involvement of smooth could be in the storage of the untranslated ribosomal protein mRNA. Mammalian hnRNP proteins are mainly located in the nucleus, although some of them have been observed to be present in the cytoplasm at low concentrations. One of these hnRNPs is the hnRNP L protein. There is also evidence indicating that some hnRNP proteins are involved in transforming/shuttling mRNA from the nucleus into the cytoplasm at a certain stage during the cell cycle (Dreyfuss and Piñol-Roma 1991; Piñol-Roma and Dreyfuss 1992). Ribosomal protein mRNA storage and translation takes place in the cytoplasm and the ribosomal protein transcripts are stored as translationally inactive post-polysomal messenger ribonucleoprotein complexes (mRNP complexes). It is not known whether the smooth protein in *Drosophila* is a nucleoprotein or whether it is actually located in the cytoplasm, but it is unlikely to have evolved a completely different function. Experimental support would be necessary to resolve this matter, which could be obtained by immunofluorescence microscopy with antibody to smooth, looking for either nucleoplasmic or cytoplasmic staining.

It would be expected, if the smooth protein acts on the ribosomal protein transcript, that the *smooth* transcripts should be present before those of the ribosomal genes. This has not been observed, since no *smooth* transcript was detected in the embryo (although there might be some transcript present) . It is therefore not really feasible, according to the expression pattern of the *smooth* transcript, that *smooth* could be involved in the storage of the ribosomal protein transcripts. Its developmental expression does not precede or correspond to that of the ribosomal protein mRNA. Although, if the protein was completely maternally contributed (and so far I have neither shown evidence for or against it), the situation would be different.

Alternatively, the smooth protein might be involved in another kind of processing of ribosomal protein pre-mRNA, rather than storage, as was suggested above. Moreover, the smooth protein might only interact with the ribosomal protein transcripts at certain stages during development, i.e. at the time the *smooth* transcripts are expressed. The *smooth* mutation might then in some way (e.g. in a similar way as in *bobbed*) retard the rate of ribosome assembly, which in turn could reduce the rate of protein synthesis, leading to the pleiotropic effects observed in the *smooth* mutants.

But whether there is any specific molecular and biochemical interaction taking place between the *smooth* gene product and the *bobbed* transcript, or the smooth protein and the *Minute* transcript has still to be shown. Therefore the main common feature between the *smooth*, the *bobbed* and *Minute* mutants is the phenotypic similarity, which might imply that they are all part of some early developmental pathway.

As a matter of interest, a mutation has been described called *abnormal abdomen* in

Drosophila mercatorum (DeSalle and Templeton 1986). Flies carrying this mutation have an abdomen with incomplete sternites and tergites, thin crinkled cuticle and bristles and hairs eliminated on the abdomen. The cause of the mutation was found to be insertions of R1 or R2 retrotransposons into the 18S or 28S rRNA, inactivating those genes.

Considering once more the possibility of the smooth protein acting in a tissue-specific fashion: there are several examples of particularly neuron-specific RRM containing proteins, one of which is the already mentioned *Drosophila* elav protein.

The elav protein is likely to be involved in neuron-specific RNA processing. Its human homologue has been discovered which is the neuronal antigen, HuD, and was found to be uniquely expressed in the brain (Szabo *et al.* 1991). The HuD protein is thought to be important in the control of the development of the human nervous system. It also shows a large degree of homology to the recently discovered *Drosophila rbp9* gene (Kim and Baker 1993b) and the sex-specific splicing protein Sxl.

Another RRM containing protein in *Drosophila*, which is thought to be essential for the PNS development in the embryo and required for adult behaviour is the couch potato protein (Bellen *et al.* 1992). This protein is thought to be involved in the PNS differentiation via alternative RNA processing. It also shows homology to Sxl, HuD and elav and is thought to regulate the RNA processing of probably more than one gene of the PNS and might show some cell specific RNA processing (Bellen *et al.* 1992).

Another indirect example of a protein which is also thought to be involved in tissue-specific splicing, in the brain and the central neurons, is the SmN protein. This small nuclear ribonucleoprotein associated polypeptide, SmN, has been suggested to add some specificity to the snRNP splicing and is thought to have some pleiotropic effects on development or functions of the nervous system. Although this protein has no RRMs, it is thought to be a splicing factor. In the human, mutations in this gene have been associated with the Prader-Willi-Syndrome (Özcelik *et al.* 1992).

Therefore a number of genes, in human and *Drosophila*, have been shown to be in particular nervous-system specific regulators involved in RNA processing. There must be a large requirement for proteins of this kind, considering the majority of genes expressed in the nervous system are alternatively spliced (Bellen *et al.* 1992).

Maybe *smooth* is involved in the splicing of nervous system-specific genes, hence the bristle effects. But *smooth* is also likely to affect other genes, since the mutant shows reduced viability and sterility effects, which are not likely to be caused by the reduction of abdominal bristles.

After all these speculations it would be interesting to find out the actual functions of smooth, but all that is known is that the smooth protein contains putative RRMs and might be involved in some way or other in the RNA processing of some transcripts. The precise role of smooth, possibly in the RNA processing mechanism, remains to be elucidated.

6.0.9 *smooth* as a quantitative trait gene

One further issue which still needs to be addressed is, what is the outcome of the *smooth* gene as a quantitative trait gene? First of all, the *smooth* gene is being considered in the context of quantitative trait genes, asking what is the nature of the *smooth* gene.

It has been suggested that quantitative trait genes might fall into groups or categories of genes representing a particular kind or type. Thompson (1975) proposed that genes affecting a quantitative character could either be control genes (e.g. enhancers or suppressors), or genes affecting direct developmental processes. As examples of the latter, the loci identified by Spickett (1966) were mentioned which affected developmental processes directly. Examples of a small number of quantitative trait genes which were identified as either regulatory or structural genes, were reported in Falconer (1989). The *bobbed* gene was mentioned as an example of a quantitative trait gene where the observed effect was more difficult to explain. In the case of *bobbed*, the bristle phenotype had been shown to be associated with mutations in the ribosomal subunits. Suggestions made, trying to explain this association, imply that *bobbed* might prevent protein biosynthesis, representing some sort of secondary effect (Falconer 1989). When comparing the situation found in *bobbed* to *smooth*, it has to be mentioned that *smooth*, like *bobbed*, does not seem to act directly on bristles. But, whereas *bobbed* codes for a ribosomal protein, *smooth* does not appear to encode such a protein.

Mackay (1992) also listed a number of suggestions on the kind of genes or about the nature of genes causing quantitative variation (see introduction).

So far, there is no concrete evidence on what the nature of quantitative trait genes could be and whether these genes can be placed into specific classes of genes. Possibly, with the rapid expansion of the usage of RFLP markers, more and more QTL will be discovered which will eventually lead to the identification of the underlying genes and hence will provide some further insight into the nature of quantitative trait loci.

As already mentioned in the introduction, some earlier selection experiments for the quantitative trait of abdominal bristles resulted in the discovery of a small number of bristle mutants.

Two of the mutants, i.e. *scute* (Yoo 1980) and *scabrous* (Jones *et al.* 1968), have been shown to be involved in the development of the nervous system, where *scute* is part of the *achaete-scute* complex which is involved in the PNS and CNS development see introduction). The *scabrous* gene, which produces a secreted protein, is thought to be involved in cellular communication during neurogenesis (Mlodzik *et al.* 1990). Both of these genes have other pleiotropic effects, apart from the bristle effects.

The third mutant was the *bobbed* mutant which, in contrast to the bristle duplications caused by the *scabrous* mutants, reduces the number of bristles (as does *smooth*) and whose phenotype was described above in greater detail. Frankham *et al.* (1978) reported it to have arisen twice independently by unequal cross-over in low selection lines.

The fact that these three mutations have been isolated as mutations in selection experiments underlines the point that new mutations do arise in response to selection, at least under laboratory conditions.

How does this compare with natural populations? Mackay and Langley (1990) asked this question when they studied molecular and phenotypic variation in the *achaete-scute* region of *Drosophila melanogaster.* In particular, they studied the effect of DNA insertions in the *achaete-scute* region of isogenic X chromosomes extracted from natural populations. Their results indicate a correlation between DNA insertions into the *achaete-scute* region and an average reduction in bristle numbers. Therefore, Mackay and Langley have shown that insertions into the AS-Complex in natural populations can cause quantitative variation in bristle number. They also argued that many of those loci affecting quantitative traits might be those which have been shown to have major qualitative effects, where the latter could be regarded as an extreme of a series of alleles. Considering these observations in the context of the isolation of the *smooth* gene, the P element insertion in the *smooth* locus seems to mimic a naturally occurring event and, to draw further parallels to Mackay and Langley's conclusion, *smooth* might be a gene of major qualitative effect with the mutation isolated showing an extreme effect.

smooth as a gene of major effect

Early this century, Nilsson-Ehle (1909 in East 1910) and East (1910) were among the first to realise that quantitative traits were governed by multiple factors of small effect following the laws of Mendelian inheritance. The factors involved in this multiple factor heredity, the genes of small effect (polygenes), were thought to be numerous and contribute collectively to the quantitative trait, modified in their effect by environmental factors. Genetic variation for quantitative characters was therefore assumed to be controlled by the segregation of multiple factors.

Due to the large number of genes all of small and indistinguishable effects, the individual alleles causing variation in quantitative traits are normally difficult to detect; and hence variation observed in quantitative characters needs to be described by means of biometrical procedures such as means, variances, heritabilities etc. (Falconer 1989).

When it came to the more detailed analysis of factors contributing to quantitative traits, it was discovered that the number of factors affecting a trait differed and that there was a range of different effects. Researchers started to dissect the factors contributing to a quantitative character and noted that quantitative trait loci differed considerably in their magnitudes of effect, hence the description "minor" and "major" genes.

Gradually, more and more evidence started to amount, indicating that for every quantitative trait there are at least a few genes of large effect among a greater number of genes of small effect. The actual number of genes influencing a quantitative trait is difficult to estimate, since a number of factors including gene frequencies, type of gene action, size of effect, linkage relationships etc. have to be considered.

The initial assumption that polygenes cannot be mapped, has proven to be untrue and an increasing number of QTLs are being mapped, showing a behaviour identical to Mendelian qualitative genes.

The *smooth* gene has been detected as a gene influencing a quantitative trait. Quantitative geneticists would regard the *smooth* gene as a gene with some mutant alleles which have a large effect on a quantitative trait, hence a major gene. *smooth* also appears to behave in a Mendelian fashion like a qualitative gene, i.e. segregating like a single autosomal recessive allele. On the other hand, *smooth* and the other four above mentioned abdominal bristle mutants, i.e. *Minute*, *bobbed*, *scute*, and *scabrous*, only make up a very small proportion of the total number of genes affecting abdominal bristle development. There are at least 145 genes affecting bristle development in some way or other (Mackay 1985), most of which also show some pleiotropic effects.

Final remarks

The initial aim of the experiment designed by Mackay (1984; 1985) had been to investigate whether P element mutagenesis could cause variation in quantitative characters in *Drosophila melanogaster*. The outcome of the experiment had been nothing less than the mapping and actual isolation of a quantitative trait gene. The approach applied to study a particular quantitative trait locus in *Drosophila* has been shown to be successful. How does this method compare to other approaches available to map QTL?

As early as in 1923, Sax (1923) started mapping a quantitative trait by making use of the genetic linkage of a different locus to the gene of interest, in this case seed size in beans. It was then shown to be possible to map QTL by cosegregation analysis, i.e. following the segregation and genetic linkage between a marker locus and the QTL.

For this purpose, a large number of genomic markers were required, preferably evenly spread over the whole length of the chromosomes. These markers, mainly morphological markers, were only available in well characterised species like maize and *Drosophila*, where early this century genetic linkage maps had already been established. A number of attempts have since been made to map QTL in *Drosophila*. The most extensive study by far was carried out by Shrimpton and Robertson (1988a; 1988b), using a recombinant approach where the chromosome under investigation was subdivided into five sections by visible marker loci and the number of genes affecting the sternopleural bristles in each of those subsections was determined. At least 17 loci affecting sternopleural bristles on the third chromosome were mapped and it was shown that an uneven distribution of these loci over the length of the chromosome existed. Until recently, this kind of classical mapping study was only feasible in organisms where a large number of morphological markers were available.

A new molecular approach to mapp QTL requires the presence of molecular linkage markers, e.g. RFLPs (Botstein *et al.* 1980), covering the whole genome of organisms. In the last few years, particular plants and animals of economic value have been marked by RFLPs which have subsequently become instrumental in establishing linkage between them and the QTL.

First associations between RFLP markers and qualitative and quantitative traits were reported by Paterson *et al.* (1988). The expectations had been that eventually the molecular characterisation of the QTL would follow, which was thought to be possible by simply walking along the chromosome, using the RFLP markers as a starting point.

As far as the genetical mapping of quantitative and qualitative genes is concerned, the RFLP markers have turned out to be very useful. But establishing the linkage between the RFLP and the marker is not sufficient for the isolation of a gene and it is important to have the RFLP very closely located, tightly linked, to the gene of interest which is segregating with the RFLP. As a matter of fact, Tanksley *et* *al.* (1992) reported recently that most RFLP maps show a spacing of more than 5 cM between each marker (which in terms of kilobase pairs equals approximately 550 kb per one cM in tomato). In order to make the molecular cloning of the genes feasible, a much larger number of markers is necessary, covering the entire genome and Tanksley *et al.* (1992) established high density linkage maps for potato and tomato, reducing the distance between the markers to 1.2 cM.

This problem of actually locating the gene of interest has not only been encountered in crop plants, but also in the study of human diseases where, although the gene of interest had been identified, it still turned out to be several thousand or even hundred thousand of kilobase pairs away from the RFLP and required extensive chromosome walking.

Therefore, the cloning of QTL by RFLP analysis has so far not led to the isolation of many genes underlying quantitative inheritance or other qualitative Mendelian genes.

For the purpose of cloning a quantitative trait gene in *Drosophila*, the approach used by Mackay (1984; 1985), making use of transposon tagging, has been successful. Due to the fact that P element mutagenesis had been used for causing mutations in a quantitative character, the P elements turned out to be an added bonus since their presence vastly facilitated not only the mapping, but also the isolation and the actual cloning of the gene.

In *Drosophila*, this approach seems to be an ideal way to study quantitative trait genes and Mackay (1992a; 1992b) has improved the experimental procedure since the original experiment, so that it is now possible to study single insertions into the *Drosophila* genome affecting any trait of interest.

Chapter 7

Appendix

7.1 Media and solutions

L-broth
10 g/l tryptone (Oxoid)
5 g/l yeast extract (Oxoid)
5 g/l NaCl

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

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

2×TY medium
16 g/l tryptone
10 g/l yeast extract
5 g/l NaCl

ψ-broth
20 g/l Bacto tryptone (Difco)
5 g/l Bacto yeast extract (Difco)
10 mM Mg^{2+}

ψ-agar
ψ-broth containing 14 g/l Bacto agar (Difco).

SOC medium
20 g/l tryptone
5 g/l yeast extract
10 mM NaCl
2.5 mM KCl
10 mM $MgCl_2$
10 mM $MgSO_4$
20 mM glucose
pH 7.0

T-broth
10 g/l tryptone (Oxoid)

5 g/l NaCl

Tris buffer
Tris-HCl (Sigma) and Tris-Base (Sigma) mixed in appropriate proportions according to the desired pH.
e.g. 1 M Tris pH 8.5:
0.7 M Tris-HCl
0.3 M Tris-Base

TE
10 mM Tris pH 8.0
1 mM EDTA pH 8.0

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

PSB
TMN made 0.05% with gelatine.

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

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

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

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

TAE
0.04 M Tris-acetate

0.002 M EDTA

20 x SSC
3 M NaCl
0.3 M *tri*sodium citrate

TfbI
30 mM potassium acetate
100 mM RbCl
10 mM $CaCl_2$
50 mM $MnCl_2$
15% by volume glycerol
Adjust to pH 5.8 with 0.2 M acetic acid and sterilised by filtration.

TfbII
10 mM PIPES
10 mM RbCl
75 mM $CaCl_2$
15% by volume glycerol
Adjust pH to 6.5 with KOH and sterilised by filtration.

STET
8% sucrose
0.5% by volume Triton X-100 (Sigma)
50 mM EDTA pH 8.0
50 mM Tris pH 8.0

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

Glucose/Tris/EDTA (GTE)
50 mM glucose
25 mM Tris-Cl pH 8.0
10 mM EDTA

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

10× Nick translation buffer (NTB)

0.5 M Tris-Cl pH 7.8
50 mM $MgCl_2$
100 mM β-mercaptoethanol

6×Ficoll Stop Buffer (FSB)

17.55% Ficoll
0.1 M Tris pH 7.5
0.1 M EDTA pH 7.5
0.1% bromophenol blue

Sequencing gel

8% Acrylamide (19:1 acrylamide:bis-acrylamide)
1.1 M Acrylamide
26 mM Bis-acrylamide
8 M urea
89 mM Tris
89 mM Boric Acid
3 mM EDTA
0.05% TEMED (v/v)
0.05% Ammonium persulfate

7.2 Primers

General primers:

M13 primers:

Forward 17mer 5' GTTTTCCCAGTCACGAC 3'

Reverse 17mer 5' CAGGAAACAGCTATGAC 3'

pBluescript II KS +/- primers:

T3 17mer 5' ATTAACCCTCACTAAAG 3'

SK 17mer 5' TCTAGAACTAGTGGATC 3'

smooth gene specific primers:

575 397⇒416 5' CCGAAAGACTGCGAATAAGC 3'

461 1204⇒1185 5' GCAAGACATCAACTGTGATG 3'

646 1203⇒1222 5' GCACAAAATCTGTCATCCAC 3'

710 1647⇒1665 5' CCTGGACCACGACACTTCC 3'

323 1677⇒1659 5' CTTATCCGTGTTGGAAGTG 3'

824 1725⇒1743 5'CTTGAAGACCAAAGAGGGC 3'

933 2045⇒2027 5' CCAATCAGTTGGTCCTCTGG 3'

916 2031⇒2049 5' GGACCAACTGATTGGAATC 3'

709 2240⇒2257 5' CCAAGAGCATGAACGGCG 3'

N.B.: The underlined numbers represent primers for the non-coding strand.

7.3 Sequences and sequence comparisons

Figure 25: cDNA sequence and translation of the *smooth* gene

Opposite, the 2613 bp of the cDNA sequence of the *smooth* gene has been listed, including the predicted translation product. The translational start site is at position 907 bp of the cDNA sequence and the stop at 2334. The numbers along either side of the sequence represent the number of basepairs.

Opposite page 194

cDNA sequence and translation

```
   1  gagtttcgaggtttcgaggttcgagctaaaactttgtgcacggagcaaattaaaaataat   60
  61  aaaataagaaataaaataataacaggcataagaaaagaaaagcgacgacaacaacaaaag   120
 121  ctgatgaatacgtgcgttcgtgtgggtgccgtgcattctgcgtacatatttacaagcatt   180
 181  cgtacgcttaaattaaatactttggcttaattaaatacttccgcggcgtcgttgttgtag   240
 241  ttgttgatattttgcctcttcacgcgttgtctacggttatcgattgtcgtcgtctcgctc   300
 301  cctcgtacgcctacgcatcgagttgttgtttgtttgttgtttgtgtctgtcctcgcgttt   360
 361  tgtttgtgtactgacaaaataacaacaaaataaaacccgaaagactgcgaataagcataa   420
 421  ataataaacaacgagacaagtgcaaaaagcgcagcgaaaaggccagctggtaaatccaaa   480
 481  tcagaggaacatccagaggaacccggaaaatttatcactccaatcacaagtgcaataaaa   540
 541  agttgaaagtgttttacgcagcgccaaagaaggaagttgaagcgaggcgaatagtaaata   600
 601  aatgagaagagcccacaacagctacaccatcatataaacaacaacccaaacatcccacat   660
 661  tcaacaccaattacaaagtaaacatttacaactcatttctcgcccaggagattcgagtgt   720
 721  gatccacagaaaacaaaaaccgcgcgcgcgtttttgaaatcgatcttcgcgcggcttcgt   780
 781  cgcacgtcacttgaatatatctttcgcgcaaccactgggctcaggcctcaaaaatccgcc   840
 841  aaaattctacaccaacaaactttaatcagccccgagatttcgctccagttcgcaacggcg   900
 901  accacaatgccctacaacggcgctagtaatggcagtggcgccagtggcgccggaggagga   960
            M  P  Y  N  G  A  S  N  G  S  G  A  S  G  A  G  G  G
 961  ggggcaaccatagtcgtcaccgagggccgcaaaacaaaaagatccgcaccggagtccag   1020
      G  A  T  I  V  V  T  E  G  P  Q  N  K  K  I  R  T  G  V  Q
1021  cagcccggggagaacgatgtgcacatgcatgctaggtccacaccacaacagaaccagcag   1080
      Q  P  G  E  N  D  V  H  M  H  A  R  S  T  P  Q  Q  N  Q  Q
1081  caagcacttatgaacaagtcaaatgacgacctacggagaaagcgtcccgagactacacgg   1140
      Q  A  L  M  N  K  S  N  D  D  L  R  R  K  R  P  E  T  T  R
1141  ccaaatcacattcttctcttcaccatcataaatcccttctatcccatcacagttgatgtc   1200
      P  N  H  I  L  L  F  T  I  I  N  P  F  Y  P  I  T  V  D  V
1201  ttgcacaaaatctgtcatccacatggccaagtgcttcgcattgtcatattcaaaaagaat   1260
      L  H  K  I  C  H  P  H  G  Q  V  L  R  I  V  I  F  K  K  N
1261  ggggtccaggccatggtcgagttcgataatctggatgcggccactagggctcgcgagaat   1320
      G  V  Q  A  M  V  E  F  D  N  L  D  A  A  T  R  A  R  E  N
1321  ctgaatggagccgacatttatgccggatgctgcactctgaaaatcgattatgccaagccg   1380
      L  N  G  A  D  I  Y  A  G  C  C  T  L  K  I  D  Y  A  K  P
1381  gagaaattgaacgtgtacaagaatgagcccgataccagctgggactatacgctgagcaca   1440
      E  K  L  N  V  Y  K  N  E  P  D  T  S  W  D  Y  T  L  S  T
1441  gaaccaccgctattgggacccggagccgctttcccaccattcggagctcccgaatatcac   1500
      E  P  P  L  L  G  P  G  A  A  F  P  P  F  G  A  P  E  Y  H
1501  accaccacaccggagaactggaagggggccgccatccatcccactggcctgatgaaggag   1560
      T  T  T  P  E  N  W  K  G  A  A  I  H  P  T  G  L  M  K  E
1561  cccgctggtgttgtgcccggacgcaatgctccggtggccttcacaccgcaaggacaggct   1620
      P  A  G  V  V  P  G  R  N  A  P  V  A  F  T  P  Q  G  Q  A
1621  cagggcgccgtcatgatggtctacggcctggaccacgacacttccaacacggataagctc   1680
      Q  G  A  V  M  M  V  Y  G  L  D  H  D  T  S  N  T  D  K  L
1681  ttcaatttggtttgcctgtacggcaacgtggcacggatcaagttcttgaagaccaaagag   1740
      F  N  L  V  C  L  Y  G  N  V  A  R  I  K  F  L  K  T  K  E
1741  ggcaccgccatggtgcaaatgggagacgctgtggccgttgagcgttgcgtgcagcacttg   1800
      G  T  A  M  V  Q  M  G  D  A  V  A  V  E  R  C  V  Q  H  L
1801  aacaacattcccgtgggcactggtggcaagatacagatcgctttctccaaacagaacttc   1860
      N  N  I  P  V  G  T  G  G  K  I  Q  I  A  F  S  K  Q  N  F
1861  ctatccgaggtgatcaacccgttcttgctgcccgatcattcgcccagcttcaaggagtac   1920
      L  S  E  V  I  N  P  F  L  L  P  D  H  S  P  S  F  K  E  Y
1921  accggctccaagaacaatcgtttcctatcgccggcccaggcgagcaagaatcgcattcaa   1980
      T  G  S  K  N  N  R  F  L  S  P  A  Q  A  S  K  N  R  I  Q
1981  ccaccgagcaagatttttgcactttttcaacacaccgcccggcttgaccgaggaccaactg   2040
      P  P  S  K  I  L  H  F  F  N  T  P  P  G  L  T  E  D  Q  L
2041  attggaatctttaacatcaaggatgtgcccgccacatcggtgcgcctgttcccctttgaag   2100
      I  G  I  F  N  I  K  D  V  P  A  T  S  V  R  L  F  P  L  K
2101  accgagcgctcgtcgtcgggactgatcgaatttccaatatctcgcaggcagtgctggct   2160
      T  E  R  S  S  S  G  L  I  E  F  S  N  I  S  Q  A  V  L  A
2161  atcatgaagtgcaaccatctgcctattgagggtaaaggcaccaagttcccattcatcatg   2220
      I  M  K  C  N  H  L  P  I  E  G  K  G  T  K  F  P  F  I  M
2221  aagctgtgcttttcctcatccaagagcatgaacggcgcctggaacaatgcggccagcgag   2280
      K  L  C  F  S  S  S  K  S  M  N  G  A  W  N  N  A  A  S  E
2281  ggcatgatcgagaaggagaacgaggtggataccaaggtggacatctacaattgagatttg   2340
      G  M  I  E  K  E  N  E  V  D  T  K  V  D  I  Y  N  *
2341  attacgattgtgtaagcaagcaaacaacaaccggcttaactagaacacaatacagaacaa   2400
2401  cactggggatatgggtcgcaaacaacgtgcagaagacaacagcaacagcaaaatcagcaa   2460
2461  caaatgcgtcaccaactgcaccaattgctgaaagttcaactacagatccactacgatcca   2520
2521  ccaacttgatggcttacacattcgacaagaaaagcgtgaatttcaaacaaattgattgct   2580
2581  aaatcaattaaacaaaccaacaaaaaaaaaaaa  2613
```

Figure 26: Frames

This figure was generated by the FRAMES program of the UWGCG package on the VAX computer. It shows all the reading frames obtainable from the cDNA sequence clone. The six horizontal lines represent the DNA sequence from 1 to 2613 bp, as indicated by the small vertical lines at the top, middle and bottom, where the scale is labelled every 500 bp.

The top three horizontal lines indicate the 5' to 3' direction of the first strand of the DNA, representing each different reading frame.
The lower three horizontal lines show the 3' to 5' direction with each possible reading frame. The vertical ticks above each horizontal DNA line indicate START codons, whereas the vertical lines below indicate STOP codons.

In this FRAMES output all START (ATG) and STOP (TAA, TGA and TAG) codons are marked.
The longest open reading frame can be seen on the first horizontal line with a START codon at position 907 bp and the STOP codon at position 2334 bp. This open reading frame corresponds to the putative *smooth* transcript.

Opposite page 195

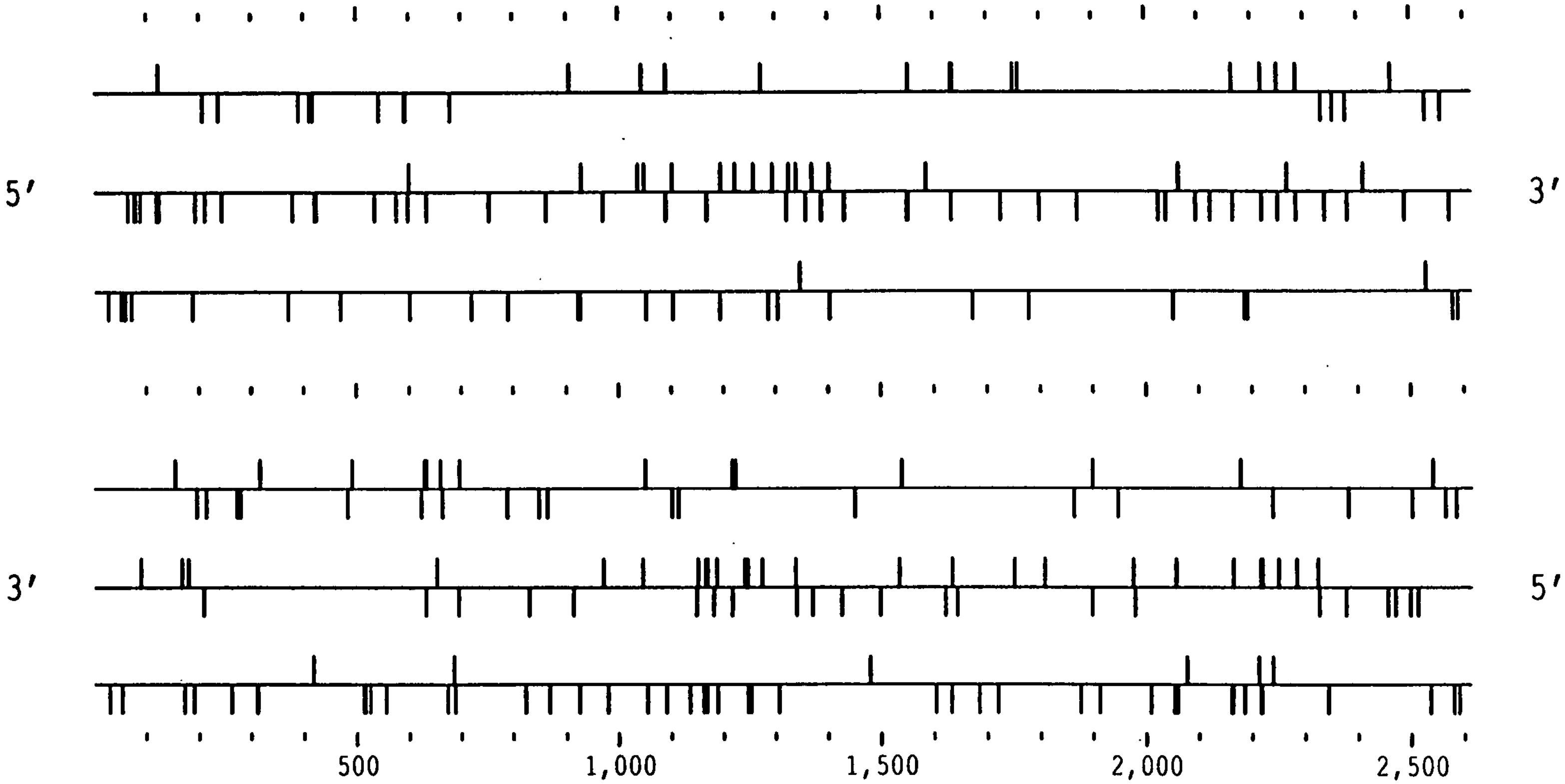

5′
3′
3′
5′
500
1,000
1,500
2,000
2,500

Figure 27: Translation of *smooth* cDNA clone from 907 bp to 2332 bp

Length: 475 amino acids

Amino acid sequence of *smooth*

```
  1  MPYNGASNGS GASGAGGGGA TIVVTEGPQN KKIRTGVQQP GENDVHMHAR

 51  STPQQNQQQA LMNKSNDDLR RKRPETTRPN HILLFTIINP FYPITVDVLH

101  KICHPHGQVL RIVIFKKNGV QAMVEFDNLD AATRARENLN GADIYAGCCT

151  LKIDYAKPEK LNVYKNEPDT SWDYTLSTEP PLLGPGAAFP PFGAPEYHTT

201  TPENWKGAAI HPTGLMKEPA GVVPGRNAPV AFTPQGQAQG AVMMVYGLDH

251  DTSNTDKLFN LVCLYGNVAR IKFLKTKEGT AMVQMGDAVA VERCVQHLNN

301  IPVGTGGKIQ IAFSKQNFLS EVINPFLLPD HSPSFKEYTG SKNNRFLSPA

351  QASKNRIQPP SKILHFFNTP PGLTEDQLIG IFNIKDVPAT SVRLFPLKTE

401  RSSSGLIEFS NISQAVLAIM KCNHLPIEGK GTKFPFIMKL CFSSSKSMNG

451  AWNNAASEGM IEKENEVDTK VDIYN
```

Figure 28: The 558 amino acid sequence of the human heterogeneous ribonucleoprotein L (Piñol-Roma *et al.* 1989))

Amino acid sequence of human heterogeneous nuclear RNP L protein

```
  1  MVKMAAAGGG GGGGRYYGGG SEGGRAPKRL KTDNAGDQHG GGGGGGGGAG
 51  AAGGGGGGEN YDDPHKTPAS PVVHIRGLID GVVEADLVEA LQEFGPISYV
101  VVMPKKRQAL VEFEDVLGAC NAVNYAADNQ IYIAGHPAFV NYSTSQKISR
151  PGDSDDSRSV NSVLLFTILN PIYSITTDVL YTICNPCGPV QRIVIFRKNG
201  VQAMVEFDSV QSAQRAKASL NGADIYSGCC TLKIEYAKPT RLNVFKNDQD
251  TWDYTNPNLS GQGDPGSNPN KRQRQPPLLG DHPAEYGGPH GGYHSHYHDE
301  GYGPPPPHYE GRRMGPPVGG HRRGPSRYGP QYGHPPPPPP PPEYGPHADS
351  PVLMVYGLDQ SKMNGDRVFN VFCLYGNVEK VKFMKSKPGA AMVEMADGYA
401  VDRAITHLNN NFMFGQKLNV CVSKQPAIMP GQSYGLEDGS CSYKDFSESR
451  NNRFSTPEQA AKNRIQHPSN VLHFFNAPLE VTEENFFEIC DELGVKRPSS
501  VKVFSGKSER SSSGLLEWES KSDALETLGF LNHYQMKNPN GPYPYTLKLC
551  FSTAQHAS
```

Figure 29: Amino acid sequence GAP comparison over the entire length of the smooth protein and the hnRNP L protein

```
GAP of: hnRNP.pep from: 1 to: 558
to: smooth.pep from: 1 to: 475

Percent Similarity: 64.640   Percent Identity: 43.919

             .         .         .         .         .
  1 MVKMAAAGGGGGGGRYYGGGSEGGRAPKRLKTDNAGDQHGGGGGGGGGAG 50
                                         ...|::.|:|:.|
  1 ....................................MPYNGASNGSGASG 14

             .         .         .         .         .
 51 AAGGGGGGENYDDPHKTPASPVVHIRGLIDGVVEADLVEALQEFGPISYV 100
    |:|||:.    ::|:... ...|: .|  |. :.|
 15 AGGGGATIVVTEGPQNKKIRTGVQQPGENDVHMHAR.............. 50

             .         .         .         .         .
101 VVMPKKRQALVEFEDVLGACNAVNYAADNQIYIAGHPAFVNYSTSQKISR 150
                           . :.:||     :.|::| |..:  .:
 51 .......................STPQQNQ.....QQALMNKSNDDLRRK 72

             .         .         .         .         .
151 PGDSDDSRSVNSVLLFTILNPIYSITTDVLYTICNPCGPVQRIVIFRKNG 200
    ..:.. .   | :|||||:||:|.||.|||..||:| |.| |||||:|||
 73 RPETTRP...NHILLFTIINPFYPITVDVLHKICHPHGQVLRIVIFKKNG 119

             .         .         .         .         .
201 VQAMVEFDSVQSAQRAKASLNGADIYSGCCTLKIEYAKPTRLNVFKNDQD 250
    ||||||||.::.| ||:..|||||||.|||||||:||||.:|||:||:.|
120 VQAMVEFDNLDAATRARENLNGADIYAGCCTLKIDYAKPEKLNVYKNEPD 169

             .         .         .         .         .
251 T.WDYTNPNLSGQGDPGSNPNKRQRQPPLLGDHPA..EYGGPHGGYHSHY 297
    | ||||   ||.             :|||||. :|  .:|:|  :||.
170 TSWDYT...LST.............EPPLLGPGAAFPPFGAP..EYHTTT 201

             .         .         .         .         .
298 HDEGYGPPPPHYEGRRMGPPVGGHRRGPSRYGPQYGHPPPPPPPPEYGPH 347
    .::  |:: .. : .: .:.|.. |.:|  :.||              .:
202 PENWKGAAIHPTGLMKEPAGVVPGRNAPVAFTPQ..............GQ 237
```

```
             .         .         .         .         .
348 ADSPVLMVYGLDQSKMNGDRVFNVFCLYGNVEKVKFMKSKPGAAMVEMAD 397
    |:::|:||||||:.. |.|::||:.||||||.::||:|.|.|.|||:|:|
238 AQGAVMMVYGLDHDTSNTDKLFNLVCLYGNVARIKFLKTKEGTAMVQMGD 287

             .         .         .         .         .
398 GYAVDRAITHLNN.NFMFGQKLNVCVSKQPAIMPG.QSYGLEDGSCSYKD 445
    : ||:|.: ||||  .  |.|:.:..|||  : .. ..: |.| |.|:|:
288 AVAVERCVQHLNNIPVGTGGKIQIAFSKQNFLSEVINPFLLPDHSPSFKE 337

             .         .         .         .         .
446 FSESRNNRFSTPEQAAKNRIQHPSNVLHFFNAPLEVTEENFFEICDELGV 495
    :.:|:|||| .|.||.|||||.||.:|||||.| ::||:.:::| :  :|
338 YTGSKNNRFLSPAQASKNRIQPPSKILHFFNTPPGLTEDQLIGIFNIKDV 387

             .         .         .         .         .
496 KRPSSVKVFSGKSERSSSGLLEWESKSDALETLGFLNHYQMKNPNGPYPY 545
     .:.||::|. |.|||||||:|:.. |:|: .:   ||..:......:|:
388 .PATSVRLFPLKTERSSSGLIEFSNISQAVLAIMKCNHLPIEGKGTKFPF 436

             .         .         .
546 TLKLCFSTAQHAS.......................... 558
    .:|||||...  .
437 IMKLCFSSSKSMNGAWNNAASEGMIEKENEVDTKVDIYN 475
```

Figure 30:

This figure (2 pages) shows six sequence comparisons. Those regions of the hnRNP L protein which show a limited homologies to the RNA-binding domains of other proteins (see Chapter 4) are depicted. The putative RNA-binding domains have here been named 'repeat domains'.

The percentage similarity and percentage identity was calculated by the GAP program of the UWGCG package.

The percentage similarity and percentage identity between these 'repeat domains' is less than the numbers obtained in a comparison between the smooth peptide and the hnRNP L peptide sequences (see figure 24).

1) hnRNP repeat domain 1 x hnRNP repeat domain 2

Percent Similarity: 44.156   Percent Identity: 23.377

```
             .         .         .         .         .
 63 DPHKTPASPVVHIRGLIDGVVEADLVEALQEFGPISYVVVMPKK.RQALV 111
             :. | . | ::... | .  .. ||:  :|::.|.  ||:|
164 ........LLFTILNPIYSITTDVLYTICNPCGPVQRIVIFRKNGVQAMV 205

             .         .         .         .         .
112 EFEDVLGACNAVNYAADNQIYIAGHPAFVNYSTSQK.............. 147
    ||:.| :| .| .  .:.:|| :. .  ::|... :
206 EFDSVQSAQRAKASLNGADIYSGCCTLKIEYAKPTRLNVFKNDQDTWDYT 255
```

2) hnRNP repeat domain 1 x hnRNP repeat domain 3

Percent Similarity: 51.389   Percent Identity: 23.611

```
             .         .         .         .         .
 63 DPHKTPASPVVHIRGLIDGVVEADLV.EALQEFGPISYVVVMPKKRQ.AL 110
             |: : || :: :::| | :.:  :| :. | .|..|.. |:
352 .........VLMVYGLDQSKMNGDRVFNVFCLYGNVEKVKFMKSKPGAAM 392

             .         .         .
111 VEFEDVLGACNAVNYAADNQIYIAGHPAFVNYSTSQK 147
    ||:.|..:. .|:.. .:| :: .     :| :.|..
393 VEMADGYAVDRAITHLNNNFMFGQK....LNVCVSKQ 425
```

3) hnRNP repeat domain 1 x hnRNP repeat domain 4

Percent Similarity: 45.283   Percent Identity: 22.642

```
             .         .         .         .         .
 63 .............DPHKTPASPVVHIRGLIDGVVEADLVEALQEFG...P 96
                 . :  .:| |:|: .   :|.|.::.|  :|:|   |
449 SRNNRFSTPEQAAKNRIQHPSNVLHFFNAPLEVTEENFFEICDELGVKRP 498

             .         .         .         .         .
 97 ISYVVVMPKKRQALVEFEDVLGACNAVNYAADNQIYIAGHPAFVNYSTSQ 146
     |  |. .|. ..  :: :
499 SSVKVFSGKSERSSSGLLE............................... 517
```

4) hnRNP repeat domain 2 x hnRNP repeat domain 3

Percent Similarity: 50.000   Percent Identity: 20.833

```
164 ..LLFTILNPIYSITTDVLYTICNPCGPVQRIVIFRKNGVQAMVEFDSVQ 211
      :::.: ..  .:..| ::.:   :| |::: :::.....||||:...
352 VLMVYGLDQS..KMNGDRVFNVFCLYGNVEKVKFMKSKPGAAMVEMADGY 399

             .         .         .         .
212 SAQRAKASLNGADIYSGCCTLKIEYAKPTRLNVFKNDQDTWDYT 255
    ..:|| . ||.. :::.            :|||  ..|
400 AVDRAITHLNNNFMFGQ............KLNVCVSKQ...... 425
```

5) hnRNP repeat domain 2x hnRNP repeat domain 4

Percent Similarity: 52.083   Percent Identity: 18.750

```
              .         .         .         .         .
164 .....................LLFTILNPIYSITTDVLYTICNPCGPVQR 192
                          :: ::|: ..:|.: ::.||:.
449 SRNNRFSTPEQAAKNRIQHPSNVLHFFNAPLEVTEENFFEICDE...... 492

             .         .         .         .         .
193 IVIFRKNGVQAMVEFDSVQSAQRAKASLNGADIYSGCCTLKIEYAKPTRL 242
    :.: |..:|..:    |..|.....: |:
493 LGVKRPSSVKVF....SGKSERSSSGLLE..................... 517
```

6) hnRNP repeat domain 3 x hnRNP repeat domain 4

Percent Similarity: 43.750   Percent Identity: 17.188

```
             .         .         .         .         .
352 ...VLMVYGLDQSKMNGDRVFNVFCLYGNVEKVKFMKSKPGAAMVEMADG 398
         :  . :..| . ::. ||: ::... .|      .:..:.|:.|:
449 SRNNRFSTPEQAAKNRIQHPSNVLHFFNAPLEV......TEENFFEICDE 492

             .         .
399 YAVDRAITHLNNNFMFGQKLNVCVSKQ.. 425
    .:|.|: .      :|:.| : : |
493 LGVKRPSSVK....VFSGKSERSSSGLLE 517
```

Figure 31:

This figure shows three sequence comparisons. Those regions of the smooth protein which show a limited homologies to the RNA-binding domains of other proteins (see Chapter 4) are depicted. The putative RNA-binding domains have here been named 'repeat domains'.

The percentage similarity and percentage identity was calculated by the GAP program of the UWGCG package.

The percentage similarity and percentage identity between these 'repeat domains' is less than the numbers obtained in a comparison between the smooth peptide and the hnRNP L peptide sequences (see figure 24).

1)

```
        smooth repeat domain 1 x smooth repeat domain 2
   Percent Similarity: 47.945   Percent Identity: 21.918

             .         .         .         .         .
 83 ..LLFTIINPFYPITVDVLHKICHPHGQVLRIVIFKKNGVQAMVEFDNLD 130
      :::.: :.  . ..| | .:.  .|.| || ::|..:. |||::::
242 VMMVYGLDHD..TSNTDKLFNLVCLYGNVARIKFLKTKEGTAMVQMGDAV 289
             .         .         .         .
131 AATRARENLNGADIYAGCCTLKIDYAKPEKLNVYKNEPDTSWDYT 175
    |..|. ::||. .: ....:.|.:.|.
290 AVERCVQHLNNIPV.GTGGKIQIAFSKQ................. 316
```

2)

```
        smooth repeat domain 1 x smooth repeat domain 3
   Percent Similarity: 38.235   Percent Identity: 22.059

             .         .         .         .         .
 83 LLFTIINPFYPITVDVLHKICHPHGQVLRIVIFKKNGVQAMVEFDNLDAA 132
       .  |.|.. . .  ::| :|.:.:|::.    |...::.|
341 ...SKNNRFLSPAQASKNRI.QPPSKILHFF....NTPPGLTE....... 375
             .         .         .         .
133 TRARENLNGADIYAGCCTLKIDYAKPEKLNVYKNEPDTSWDYT 175
              | . |. .:|   |.. :| ..|.|...|  ..
376 ..........DQLIGIFNIKDVPATSVRLFPLKTERSSSGLIE 408
```

3)

```
        smooth repeat domain 2 x smooth repeat domain 3
   Percent Similarity: 32.203   Percent Identity: 15.254
             .         .         .         .         .
242 .........VMMVYGLDHDTSNTDKLFNLVCLYGNVARIKFLKTKEGTAM 282
             .    .  :..|.. .:|| .. ..: . | ::..|:..|
341 SKNNRFLSPAQASKNRIQPPSKILHFFNTPPGLTEDQLIGIFNIKDVPAT 390
             .         .         .
283 VQMGDAVAVERCVQHLNNIPVGTGGKIQIAFSKQ 316
      .  :: .||:   | :
391 SVRLFPLKTERSSSGLIE................ 408
```

7.4 Acknowledgments

I wish to thank my supervisor Dr Andrew Leigh Brown for his support and advice and the careful reading of my thesis. In addition, I would also like to thank a series of second supervisors, in particular Dr John Bishop and Dr David Finnegan.

I am grateful to a number of people and groups of the former Genetics Department: James Macpherson for the help during my first year in the lab; past and present members of the HIV group, in particular Peter Balfe; Sarah Ashelford, who helped me with the *in vitro* translation of *smooth*; the Protozoan Group for the use of their Macintosh computer; the West Building for the lunch-time entertainment; Ian Hastings for helpful suggestions in passing; Herbert Sauro (mainly on-line via email) and Rankin Small for answering all my computing queries; Graham Brown and Frank Johnston for their photographic service.

But most of all I would like to thank Sarah Ross, not only for the salivary gland *in situ* hybridisations she carried out for me, but also for her company in the lab, hundreds of cups of tea and the odd multiple of 12 minipreps she prepared for me!

I would like to thank Marcel Wehrli for the gift an aliquot of Nick Brown's cDNA library.

Bibliography

[1] S.A. Adam, T. Nakagawa, M.S. Swanson, T.K. Woodruff, and G. Dreyfuss. mRNA polyadenylate-binding protein: gene isolation and sequencing and identification of a ribonucleoprotein consensus sequence. *Mol. Cell. Biol.*, 6:2932–2943, 1986.

[2] H. Amrein, T. Maniatis, and R. Nöthiger. Alternatively spliced transcripts of the sex-determining gene *tra-2* of *Drosophila* encode functional proteins of different size. *EMBO J.*, 9:3619–3629, 90.

[3] M. Ashburner. *Drosophila A laboratory handbook.* Cold Spring Harbor Laboratory Press, 1989.

[4] E.E. Ball, E.J. Rehm, and C.S. Goodman. Cloning of a grasshopper cDNA coding for a protein homologous to the A1, A2/B1 proteins of mammalian hnRNP. *Nucleic Acids Research*, 19:397, 1991.

[5] D.W. Ballard, W.M. Philbrick, and A.L.M. Bothwell. Identification of a novel 9-kDa polypeptide from nuclear extracts - DNA-binding properties, primary structure, and *in vitro* expression. *J. Biol. Chem.*, 263:8450–8457, 1988.

[6] R.J. Bandziulis, M.S. Swanson, and G. Dreyfuss. RNA-binding proteins as developmental regulators. *Genes Dev.*, 3:431–437, 1989.

[7] A.G. Bang and J.W. Posakony. The *Drosophila* gene *Hairless* encodes a novel basic protein that controls alternative cell fates in adult sensory organ development. *Genes Dev.*, 6:1752–1769, 1992.

[8] J.D. Beggs, V.D. Berg, A.V. Ooyen, and C. Weissman. Abnormal expression of a chromosomal rabbit β-globin gene in *Saccharomyces cerevisiae. Nature*, 283:835–840, 1980.

[9] L.R. Bell, E.M. Maine, P. Schedl, and T.W. Cline. *Sex-lethal*, a Drosophila sex determination switch gene, exhibits sex-specific RNA splicing and sequence similarity to RNA binding proteins. *Cell*, 55:1037–1046, 1988.

[10] H.J. Bellen, S. Kooyer, D. D'Evelyn, and J. Pearlman. The *Drosophila* couch potato protein is expressed in nuclei of peripheral neuronal precursors and shows homology to RNA-binding proteins. *Genes Dev.*, 6:2125–2136, 1992.

[11] H.J. Bellen, C.J. O'Kane, C. Wilson, R.K. Pearson U. Grossniklaus, and W.J. Gehring. P-element-mediated enhancer detection: A versatile method to study development in *Drosophila. Genes Dev.*, 3:1288–1300, 1989.

[12] M. Bennett, S. Michaud, J. Kingston, and R. Reed. Protein components specifically associated with prespliceosome and spliceosome complexes. *Genes Dev.*, 6:1986–2000, 1992a.

[13] M. Bennett, S. Pinõl-Roma, D. Staknis, G. Dreyfuss, and R. Reed. Differential binding of heterogeneous nuclear ribonucleoproteins to mRNA precursors proior to spliceosome assembly *in vitro. Mol. Cell. Biol.*, 12:3165–3175, 1992b.

[14] E. Bier, H. Vaessin, S. Shepherd, K. Lee, K. McCall, S. Barbel, L. Ackerman, R. Carretto, T. Uemura, E. Grell, L.Y. Jan, and Y.N. Jan. Searching for pattern and mutation in the *Drosophila* genome with a p-*lacZ* vector. *Genes Dev.*, 3:1273–1287, 1989.

[15] P.M. Bingham, M.G. Kidwell, and G.M. Rubin. The molecular basis of P-M hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family. *Cell*, 29:995–1004, 1982.

[16] P.M. Bingham, R. Levis, and G.M. Rubin. The cloning of the DNA sequences from the *white* locus of *Drosophila melanogaster* using a novel and general method. *Cell*, 25:693–703, 1981.

[17] H.C. Birnboim and J. Doly. A rapid alkaline extraction method for screening recombination plasmid DNA. *Nucleic Acids Research.*, 7:1513–1523, 1979.

[18] D.M. Black, M.S. Jackson, M. G. Kidwell, and G.A. Dover. KP elements repress P-induced hybrid dysgenesis in *Drosophila melanogaster. EMBO J.*, 6:4125–4135, 1987.

[19] A.L.M. Bothwell, D.W. Ballard, W.M. Philbrick, G. Lindwall, S.E. Maher, M.M. Bridgett, S.F. Jamison, and M.A. Garcia-Blanco. Murine polypyrimidine tract binding protein. *J. Biol. Chem.*, 266:24657–24663, 1991.

[20] D. Botstein, R.L. White, M. Skolnick, and R.W. Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphism. *Am. J. Hum. Genet.*, 32:314–331, 1980.

[21] E.L. Breese and K. Mather. The organisation of polygenic activity within a chromosome in Drosophila. 1. Hair characters. *Heredity*, 11:373–395, 1957.

[22] V. Brendel and S. Karlin. Association of charge clusters with functional domains of cellular transcription factors. *PNAS*, 86:5698–5702, 1989.

[23] C.B. Bridges and K.S. Brehme. *The mutants of Drosophila melanogaster.* Carnegie Inst. Wash. Publ. no.522. Washington D.C., 1944.

[24] N.H. Brown and F.C. Kafatos. Functional cDNA libraries from *Drosophila* embryos. *J. Mol. Biol.*, 203:425–437, 1988.

[25] N.H. Brown, D.L. King, M. Wilcox, and F.C. Kafatos. Developmentally regulated alternative splicing of Drosophila integrin PS2 α transcripts. *Cell*, 59:185–195, 1989.

[26] F. Brunel, P.M. Alzari, P. Ferrara, and M.M. Zakin. Cloning and sequencing of PYBP, a pyrimidine-rich specific single strand DNA-binding protein. *Nucleic Acids Research*, 19:5237–5245, 1991.

[27] P.J. Bryant. Pattern formation in imaginal discs, in The genetics and biology of *Drosophila* , ed. by M. Ashburner and T.R.F. Wraight, volume 2c. Academic Press Inc., 1978.

[28] E. Buerki, C. Anjard, J.-C. Scholdern, and C.D. Reymond. Isolation of two genes encdoing putative protein kinases regulated during *Dictyostelium discoideum* development. *Gene*, 102:57–65, 1991.

[29] M. Buvoli, F.Cobianchi, G. Biamonti, and S.Riva. Recombinant hnRNP protein A1 and its N-terminal domain show preferential affinity for oligodeoxynucleotides homologous to intron/exon acceptor sites. *Nucleic Acids Research*, 18:6595–6600, 1990.

[30] A. Caballero, M.A. Toro, and C. Lòpez-Fanjul. The response to artificial selection from new mutations in *Drosophila melanogaster. Genetics*, 127:89–102, 1991.

[31] S. Campuzano and J. Modolell. Patterning of the *Drosophila* nervous system: the *achaete-scute* gene complex. *TIG*, 8:202–208, 1992.

[32] D.R. Cavener. Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates. *Nucleic Acids Research*, 15:1353–1361, 1987.

[33] D.R. Cavener and S.C. Ray. Eukaryotic start and stop translation sites. *Nucleic Acids Research*, 19:3185–3192, 1991.

[34] B. Charlesworth and C.H. Langley. The population genetics of *Drosophila* transposable elements. *Ann.Rev.Genet.*, 23:251–287, 1989.

[35] W. Chia, G. Howes, M. Martin, Y.B. Meng, K. Moses, and S. Tsubota. Molecular anaysis of the *yellow* locus of *Drosophila. EMBO J.*, 5:3597–3605, 1986.

[36] Y.D. Choi and G. Dreyfuss. Isolation of the heterogeneous nuclear ribonucleoprotein complex (hnRNP): A unique supramolecular assembly. *PNAS*, 81:7471–7475, 1984.

[37] P. Chomczynski and N. Sacci. Single–step method of RNA isolation by acid guanidium thiocyanate–phenol–chloroform extraction. *Analy. Biochem.*, 162:156–159, 1987.

[38] T.-B. Chou, Z. Zachar, and P.M. Bingham. Developmental expression of a regulatory gene is programmed at the level of splicing. *EMBO J.*, 6:4095–4104, 1987.

[39] L. Clarke and J. Carbon. A colony bank containing synthetic Col E1 hybrid plasmids representative of the entire *E. coli* genome. *Cell*, 9:91–99, 1976.

[40] J.H. Claxton. Some quantitative features of *Drosophila* sternite bristle patterns. *Aust. J. Biol. Sci.*, 27:533–543, 1974.

[41] G.A. Clayton, G.R. Knight, J.A. Morris, and A. Robertson. An experimental check on quantitative genetical theory. III. Correlated responses. *J. Genetics*, 55:171–180, 1957.

[42] G.A. Clayton, J.A. Morris, and A. Robertson. An experimental check on quantitative genetical theory. I. Short-term responses to selection. *J. Genetics*, 55:131–151, 1957.

[43] G.A. Clayton and A. Robertson. Mutation and quantitative variation. *American Naturalist*, *LXXXIX*:151–158, 1955.

[44] G.A. Clayton and A. Robertson. An experimental check on quantitative genetical theory. II. The long-term effects of selection. *J.Genetics*, 55:152–170, 1957.

[45] F. Cobianchi, R.L. Karpel, K.R. Williams, V. Notario, and S.H. Wilson. Mammalian heterogeneous nuclear ribonucleoprotein complex protein A1. *J. Biol. Chem.*, 263:1063–1071, 1988.

[46] L. Cooley, R. Kelley, and A. Spradling. Insertional mutagenesis of the *Drosophila* genome with single P elements. *Science*, 239:1121–1128, 1988.

[47] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt. *A model of evolutionary change in proteins, in Atlas of protein sequence and structure, edited by M.O. Dayhoff*, volume 5. NBRF, Washington, 1978.

[48] R. DeSalle and A.R. Templeton. The molecular through ecological genetics of abnormal abdomen. III. Tissue-specific differential replication of ribosomal genes modulates the abnormal abdomen phenotype in *Drosophila mercatorum*. *Genetics*, 112:877–886, 1986.

[49] J. Devereux, H. Haeberli, and O. Smithies. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.*, 12:387–395, 1984.

[50] R.C. Dickson, J. Hopper, L. Mylin, and C.J. Gerardot. Sequence conservation in the *Saccharomyces* and *Kluveromyces* GAL1 transcription activators suggests functional domains. *Nucleic Acids Research*, 19:5345–5350, 1991.

[51] J. Doebley and A. Stec. Genetic-analysis of the morphological differences between maize and teosinte. *Genetics*, 129:285–295, 1991.

[52] G. Dreyfuss, S.A. Adam, and Y.D. Choi. Physical change in cytoplasmic messenger ribonucleoproteins in cells treated with inhibitors of mRNA transcription. *Mol.Cell.Biol.*, 4:415–423, 1984a.

[53] G. Dreyfuss, Y.D. Choi, and S.A. Adam. Characterization of hnRNA-protein complexes *in vivo* with monoclonal antibodies. *Mol.Cell.Biol.*, 4:1104–1114, 1984b.

[54] G. Dreyfuss, M.S. Swanson, and S. Piñol-Roma. Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation. *1988*, 13:TIBS, 1988.

[55] W.F. Eanes, C. Wesley, J. Hey, D. Houle, and J.W. Ajioka. The fitness consequences of P element insertion in *Drosophila melanogaster*. *Genet. Res.*, 52:17–26, 1988.

[56] E.M. East. A Mendelian interpretation of variation that is apparently continuous. *American Naturalist*, *XLIV*:65–82, 1910.

[57] E.M. East. Studies on size inheritance in Nicotiana. *Genetics*, 1:164–176, 1916.

[58] M.D. Edwards, C.W. Stuber, and J.F. Wendel. Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. numbers, genomic distribution and types of gene action. *Genetics*, 116:113–125, 1987.

[59] R. Eker. The recessive mutant *engrailed* in *Drosophila melanogaster. Hereditas*, 12:217–222, 1929.

[60] W.R. Engels. Hybrid dysgenesis in *Drosophila melanogaster*: rules of inheritance of female sterility. *Genet.Res.*, 33:219–236, 1979.

[61] W.R. Engels. P elements in *Drosophila. in Mobile DNA, edited by D.E.Berg and M.M.Howe*, American Society for Microbiology, Washington, D.C., 1989.

[62] W.R. Engels. The origin of P elements in *Drosophila melanogaster. BioEssays*, 14:681–686, 1992.

[63] W.R. Engels, D.M. Johnson-Schlitz, W.B.Eggleston, and J.Sved. High-frequency P element loss in *Drosophila* is homolog dependent. *Cell*, 62:515–525, 1990.

[64] D.S. Falconer. *Introduction to Quantitative Genetics.* Longman Scientific and Technical, third edition, 1989.

[65] D.J. Finnegan. Eukaryotic transposable elements and genome evolution. *TIG*, 5 no.4:103–107, 1989.

[66] D.J. Finnegan and D.H.Fawcett. *Transposable elements in Drosophila melanogaster.*, volume 3. Oxf. Surv. Eukaryotic Genes, 1986.

[67] B.J. Fitzpatrick and J.A. Sved. High levels of fitness modifiers induced by hybrid dysgenesis in *Drosophila melanogaster. Genet. Res. Camb.*, 48:89–94, 1986.

[68] R. Frankham, D.A. Briscoe, and R.K. Nurthen. Unequal crossing over at the rRNA locus as a source of quantitative genetic variation. *Nature*, 272:80–81, 1978.

[69] R. Frankham and R.K. Nurthen. Forging links between population and quantitative genetics. *Theor.Appl.Genet.*, 59:251–263, 1981.

[70] A. Garcia-Bellido and J.R. Merriam. Clonal parameters of tergite development in *Drosophila. Dev. Biol.*, 26:264–276, 1971.

[71] M.A. Garcia-Blanco, S.F. Jamison, and P.A. Sharp. Identification and purification of a 62,000-dalton protein that binds specifically to the polypyrimidine tract of introns. *Genes Dev.*, 3:1874–1886, 1989.

[72] A. Ghetti, S. Piñol-Roma, W.M. Michael, C. Morandi, and G. Dreyfuss. hnRNP I, the polypyrimidine tract-binding protein: distinct nuclear localization and association with hnRNAs. *Nucleic Acids Research*, 20:3671–3678, 1992.

[73] A. Ghysen and C. Dambly-Chaudière. Genesis of the *Drosophila* peripheral nervous system. *TIG*, 5:251–255, 1989.

[74] A. Gil, P.A. Sharp, S.F. Jamison, and M.A. Garcia-Blanco. Characterization of cDNAs encoding the polypyrimidine tract-binding protein. *Genes Dev.*, 5:1224–1236, 1991.

[75] G.B. Gloor, N.A. Nassif, D.M. Johnson-Schlitz, C.R. Preston, and W.R. Engels. Targeted gene replacement in *Drosophila* via P element-induced gap repair. *Science*, 253:1110–1117, 1991.

[76] T.J. Goralski, J.-E. Edström, and B.S. Baker. The sex determination locus *transformer-2* of *Drosophila* encodes a polypepetide with similarity to RNA binding proteins. *Cell*, 43:603–613, 1989.

[77] M. Görlach, M. Wittekind, R.A. Beckman, L. Mueller, and G. Dreyfuss. Interaction of the RNA-binding domain of the hnRNP C proteins with RNA. *EMBO J.*, 11:3289–3295, 1992.

[78] M.M. Green. The genetics of a mutable gene at the white locus of *Drosophila melanogaster. Genetics*, 56:467–482, 1967.

[79] M.M. Green. Controlling element mediated transpositions of the *white* gene in *Drosophila melanogaster. Genetics*, 61:429–441, 1969.

[80] M.M. Green. *Eukaryotic transposable elements as mutagenic agents edited by M.E. Lambert and J.F. McDonald and I.B. Weinstein*, volume 30. Banbury Report, 1988.

[81] E. Gründl and L. Dempfle. Effects of spontaneous and induced mutations on selection response. *Proc. 4th World Congress on Genetics, XIII*:177–184, 1990.

[82] C. Guthrie and B. Patterson. Splicosomal snRNAs. *Annu. Rev. Genet.*, 22:387–419, 1988.

[83] D. Hanahan. Studies on transformation of *Escherichia coli* with plasmids. *J. Mol. Biol.*, 166:557–580, 1983.

[84] D. Hanahan and M. Meselson. Plasmid screening at high colony density. *Gene*, 10:63–67, 1980.

[85] D.S. Harper, L.D. Fresco, and J.D. Keene. RNA binding specificity of a *Drosophila* snRNP protein that shares sequence homolgy with mammalian U1-A and U"-B" proteins. *Nucleic Acids Research*, 20:3645–3650, 1992.

[86] V. Hartenstein and J.W. Posakony. Development of adult sensilla on the wing and notum of *Drosophila melanogaster*. *Development*, 107:389–405, 1989.

[87] S.R. Haynes, D. Johnson, G. Raychaudhuri, and A.L. Beyer. The *Drosophila Hrb87F* gene encodes a new member of the A and B hnRNP protein group. *Nucleic Acids Research*, 19:25–31, 1991.

[88] S.R. Haynes, G. Raychaudhuri, and A.L. Beyer. The *Drosophila Hrb98DE* locus encodes four protein isoforms homologous to the A1 protein of mammalian heterogeneous nuclear ribonucleoprotein complexes. *Mol. Cell. Biol.*, 10:316–323, 1990.

[89] S.R. Haynes, M.L. Rebbert, B.A. Mozer, F. Forquignon, and I.B. Dawid. *pen* repeat sequences are GGN clusters and encode a glycine-rich domain in a *Drosophila* cDNA homologous to the rat helix destabilizing protein. *PNAS*, 84:1819–1823, 1987.

[90] M.L. Hedley and T. Maniatis. Sex-specific splicing and polyadenylation of *dsx* pre-mRNA requires a sequence that binds specifically to tra-2 protein *in vitro*. *Cell*, 65:579–586, 1991.

[91] W.G. Hill. Rates of change in quantitative traits from fixation of new mutations. *PNAS*, 79:142–145, 1982a.

[92] W.G. Hill. Predictions of response to artificial selection from new mutations. *Genet. Res. Camb.*, 40:255–278, 1982b.

[93] W.G. Hill. *Quantitative Genetics, PartII: Selection, in Benchmark Papers in Genetics Series*. Van Nostrand Reinhold, 1984.

[94] Y. Hiraizumi. Spontaneous recombination in *Drosophila melanogaster* males. *PNAS*, 68:268–270, 1971.

[95] D.W. Hoffman, C.C. Query, B.L. Golden, S.W. White, and J.D.Keene. RNA-binding domain of the A protein component of the U1 small nuclear ribonucleoprotein analyzed by NMR spectroscopy is structurally similar to ribosomal proteins. *PNAS*, 88:2495–2499, 1991.

[96] B. Hohn. *In vitro* packagaing of λ and cosmid DNA. *Methods in Enzymology*, 68:299–309, 1979.

[97] B. Hollingdale. Analyses of some genes from abdominal bristle number selection lines in *Drosophila melanogaster. Theor. Appl. Genetics*, 41:292–301, 1971.

[98] B. Hollingdale and J.S.F. Barker. Selection for increased abdominal bristle number in *Drosophila melanogaster. Theor. App. Genetics*, 41:208–215, 1971.

[99] D.S. Holmes and M. Quigley. A rapid boiling method for the preparation of bacterial plasmids. *Analytical Biochem.*, 114:193–197, 1981.

[100] T. Hunt. False starts in translational control of gene expression. *Nature*, 316:580–581, 1985.

[101] P.W. Ingham, S.M. Pinchin, K.R. Howard, and D. Ish-Horowicz. Genetic anaysis of the *hairy* locus in *Drosophila melanogaster. Genetics*, 111:463–486, 1985.

[102] K. Inoue, K. Hoshijima, H. Sakamoto, and Y. Shimura. Binding of the *Drosophila sex-lethal* gene product to the alternative splice site of *transformer* primary transcript. *Nature*, 344:461–463, 1990.

[103] M. Iwasaki, K. Okumura, Y. Kondo, T. Tanaka, and H. Igarashi. cDNA cloning of a novel heterogeneous nuclear ribonucleoprotein gene homologue in *Caenorhabditis elegans* using hamster prion protein cDNA as a hybridization probe. *Nucleic Acids Reseach*, 20:4001–4007, 1992.

[104] Y.N. Jan and L.Y. Jan. Genes required for specifying cell fates in Drosophila embryonic sensory nervous system. *TINS*, 13:493–498, 1990.

[105] S.K. Jang, M.V. Davies, R.J. Kaufman, and E. Wimmer. Initiation of protein synthesis by internal entry of ribosomes into the 5' nontranslated region of encephalomyocarditis virus RNA *in vivo. J. Virol.*, 63:1651–1660, 1989.

[106] T.-H. Jessen, C. Oubridge, C.H. Teo, C. Pritchard, and K. Nagai. Identification of molecular contacts betwen the U1 A small nuclear ribonucleoprotein and U1 RNA. *EMBO J.*, 10:3447–3456, 1991.

[107] D. St Johnston, D. Beuchle, and C. Nüsslein-Volhard. *staufen*, a gene required to localize maternal RNAs in the *Drosophila* egg. *Cell*, 66:51–63, 1991.

[108] L.P. Jones, R. Frankham, and J.S.F. Barker. The effects of population size and selection intensity in selection for a quantitative character in *Drosophila*. *Genet. Res.*, 12:249–266, 1968.

[109] L. Kalfayan and P.C. Wensink. Developmental regulation of *Drosophila* α tubulin genes. *Cell*, 29:91–98, 1982.

[110] R.E. Karess and G.M. Rubin. Analysis of P transposable element functions in *Drosophila*. *Cell*, 38:135–146, 1984.

[111] J. Karn, S. Brenner, L. Barnett, and G. Cesareni. Novel bacteriophage λ cloning vector. *PNAS*, 77:5172, 1980.

[112] P.D. Kaufman and D.C. Rio. P element transposition *in vitro* proceeds by a cut-and-paste mechanism and uses GTP as a cofactor. *Cell*, 69:27–39, 1992.

[113] P.D. Kaufmann, R.F. Doll, and D.C. Rio. *Drosophila* P element transposase recognizes internal P element DNA sequences. *Cell*, 59:359–371, 1989.

[114] P.D. Kaufmann and D.R. Rio. *Drosophila* P element transposase acts as a transcriptional repressor *in vitro*. *PNAS*, 88:2613–2617, 1991.

[115] B.K. Kay, R.K. Sawhney, and S.H. Wilson. Potential for 2 isoforms of the A1-ribonucleoprotein in *Xenopus laevis*. *PNAS*, 87:1367–1371, 1990.

[116] M.A. Kay and M. Jacobs-Lorena. Developmental genetics of ribosome synthesis in *Drosophila*. *TIGS*, 3:347–351, 1987.

[117] M.R. Kelley, S. Kidd, R.L. Berg, and M.W. Young. Restricition of P-element insertions at the *notch* locus of D*rosophila melanogaster*. *Mol.Cell Biol.*, 7:1545–1548, 1987.

[118] D.J. Kenan, C.C. Query, and J.D. Keene. RNA recognition: towards identifying determinants of specificity. *TIBS*, 16:214–220, 1991.

[119] M.G. Kidwell and J.F. Kidwell. Cytoplasm-chromosome interactions in *Drosophila melanogaster*. *Nature*, 253:755–756, 1975.

[120] M.G. Kidwell and J.F. Kidwell. Selection for male recombination in *Drosophila melanogaster. Genetics*, 84:333–351, 1976.

[121] M.G. Kidwell, J.F. Kidwell, and J.A. Sved. Hybrid dysgenesis in *Drosophila melanogaster*: a syndrome of aberrant traits including mutation, sterility and male recombination. *Genetics*, 86:813–833, 1977.

[122] M.J. Kidwell. *P-M mutagenesis,in Drosophila, a practical approach, ed. D.B.Roberts.* IRL Press, Oxford, 1986.

[123] M. Kiledjian and G. Dreyfuss. Primary structure and binding activivty of the hnRNP U protein: binding RNA through RGG box. *EMBO J.*, 11:2655–2664, 1992.

[124] Y.-J. Kim and B.S. Baker. Isolation of RRM-type RNA-binding protein genes and the analysis of their relatedness by using a numerical approach. *Mol. Cell. Biol.*, 13:174–183, 1993a.

[125] Y.-J. Kim and B.S. Baker. The *Drosophila* gene *rbp9* encodes a protein that is a member of a conserved group of putative RNA binding proteins that are nervous system-specific in both flies and humans. *J.Neurosci.*, 13:1045–1056, 1993b.

[126] Y.-J. Kim, P. Zuo, J.L. Manley, and B.S. Baker. The *Drosophila* RNA-binding protein RBP1 is localized to transcriptionally active sites of chromosomes and shows a functional similarity to human splicing factor ASF/SF2. *Genes Dev.*, 1992:2569–2579, 1992.

[127] O. Kitagawa. The effects of X-ray irradiaation on selection response in *Drosophila melanogaster. Japan. J. Genetics*, 42:121–137, 1967.

[128] N. Kleckner. Transposable elements in prokaryotes. *Ann. Rev. Gen.*, 15:341–404, 1981.

[129] K. Kongsuwan, Q. Yu, A. Vincent, M.C. Frisardi, M. Rosbach, J.A. Lengyel, and J. Merriam. A *Drosophila Minute* gene encodes a ribosomal protein. *Nature*, 317:555–558, 1985.

[130] K. Kornfeld, R.B. Saint, P.A. Beachy, H.J. Harte, D.A. Peattie, and D.S. Hogness. Structure and expression of a family of *Ultrabithorax* mRNAs generated by alternative splicing and polyadenylation in *Drosophila. Genes Dev.*, 3:243–258, 1989.

[131] M. Kozak. Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. *Microbiol. Reviews*, 47:1–45, 1983.

[132] M. Kozak. The scanning model for translation: an update. *J. Cell. Biol.*, 108:229–241, 1989.

[133] M. Kozak. An analysis of vertebrate mRNA sequences: intimations of translational control. *J. Cell Biol.*, 115:887–903, 1991.

[134] A. Kumar, C.J. Luneau J.R. Casas-Finet, R.L. Karpel, B.M. Merrill, K.R. Williams, and S.H. Wilson. Mammalian heterogeneous nuclear ribonucleoprotein A1. *J. Biol. Chem.*, 265, 1990.

[135] A. Kumar, K.R. Williams, and W. Szer. Purification and domain structure of core hnRNP proteins A1 and A" and their relationship to single-stranded DNA-binding proteins. *J. Biol. Chem.*, 261:11266–11273, 1986.

[136] U.K. Laemmli. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, 227:680–685, 1970.

[137] C. Lai and T.F.C. Mackay. Hybrid dysgenesis-induced quantitative variation on the X chromosome of *Drosophila melanogaster*. *Genetics*, 124:627–636, 1990.

[138] C.D. Laird and B.J. McCarthy. Nucleotide sequence homology within the genome of *Drosophila melanogaster*. *Genetics*, 60:323–334, 1968.

[139] C.H. Langley, A.E. Shrimpton, T. Yamazaki, N. Miyashita, Y. Matsuo, and C.F.Aquadro. Naturally occurring variation in the restriction map of the *Amy* region of *Drosophila melanogaster*. *Genetics*, 119:619–629, 1988.

[140] V. Lantz, L. Ambrosio, and P. Schedl. The *Drosophila orb* gene is predicted to encode sex-specific germline RNA-binding proteins and has localized transcripts in ovaries and early embryos. *Development*, 115:75–88, 1992.

[141] F.A. Laski, D.C. Rio, and G.M. Rubin. Tissue specificity of Drosophila P element transposition is regulated at the level of mRNA splicing. *Cell*, 44:7–19, 1986.

[142] A. Laughon, A.M. Boulet, J.R. Bermingham, R.A. Laymon, and M.P. Scott. Structure of transcripts from the homeotic *Antennapedia* gene of *Drosophila melanogaster*: Two promoters control the major protein-coding region. *Mol. Cell. Biol.*, 1986:4676–4689, 1986.

[143] B. Lemaitre and D. Coen. P regulatory products repress *in vivo* the P promoter activity in P-*lacZ* fusion genes. *PNAS*, 88:4419–4423, 1991.

[144] A.J. Leigh Brown and J.E. Moss. Transposition of the *I* element and *copia* in a natural population of *Drosophila melanogaster. Genet. Res.*, 49:121–128, 1987.

[145] R. Levis, K. O'Hare, and G.M. Rubin. Effects of transposable element insertions on RNA encoded by the *white* gene of *Drosophila. Cell*, 38:471–481, 1984.

[146] D.L. Lindsley and E.H. Grell. *Genetic variations of Drosophila melanogaster.* Carnegie Inst.Wash.Publ.627, 1968.

[147] D.L. Lindsley and G. Zimm. The genome of *Drosophila melanogaster. Droso. Inform. Serv.*, 68:215, 1990.

[148] D.L. Lindsley and G.G. Zimm. *The genome of* Drosophila melanogaster. Academic Press Inc., Harcourt Brace Jovanich Publishers, 1992.

[149] M.D. Ludevid, M.A. Freire, J.Gómez amd C.G. Burd, F. Albericio, E. Giralt, G. Dreyfuss, and M. Pagés. RNA binding characteristics of a 16 kDa glycine-rich protein from maize. *The Plant J.*, 2:999–1003, 1992.

[150] M. Lynch. The rate of polygenic mutation. *Genet. Res.*, 51:137–148, 1988.

[151] T.F.C. Mackay. Jumping genes meet abdominal bristles: Hybrid dysgenesis-induced quantitative variation in *Drosophila melanogaster. Genet.Res.*, 44:231–237, 1984.

[152] T.F.C. Mackay. Transposable element-induced response to artificial selection in *Drosophila melanogaster. Genetics*, 111:351–374, 1985.

[153] T.F.C. Mackay. Transposable element-induced fitness mutations in *Drosopila melanogaster. Genet.Res.*, 48:77–87, 1986.

[154] T.F.C. Mackay. Transposable element-induced polygenic mutations in *Drosophila melanogaster. Genet.Res.*, 49:225–233, 1987.

[155] T.F.C. Mackay. Transposable elements and fitness in *Drosophila melanogaster. Genome*, 31:284–295, 1989a.

[156] T.F.C. Mackay. Mutation and the origin of quantitative variation. *in Evolution and Animal Breeding, edited by W.G. Hill and T.F.C. Mackay.*, C.A.B. International, Wallingford.:113–119, 1989b.

[157] T.F.C. Mackay and C.H. Langley. Molecular and phenotypic variation in the *achaete-scute* region of *Drosophila melanogaster. Nature*, 348:64–66, 1990.

[158] T.F.C. Mackay, R. Lyman, M.S. Jackson, C. Terzian, and W. G. Hill. Polygenic mutation in *Drosophila melanogaster*: estimates from divergence among inbred strains. *Evolution*, 46:300–316, 1992b.

[159] T.F.C. Mackay, R.F. Lyman, and M.S. Jackson. Effects of P element insertions on quantitative traits in *Drosphila melanogaster. Genetics*, 130:315–332, 1992a.

[160] M.M. Madhavan and K. Madhavan. Morphogenesis of the epidermis of adult abdomen of *Drosophila. J. Embryol. exp. Morph.*, 60:1–31, 1980.

[161] W.H. Mager. Control of ribosomal protein gene expression. *BBA*, 949:1–15, 1987.

[162] T. Maniatis, E.F. Fritsch, and J. Sambrook. *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor., 1982.

[163] K. Mather. Polygenic inheritance and natural selection. *Biol. Rev. Cam. Phil. Soc.*, 18:32–64, 1943.

[164] E.L. Matunis, M.J. Matunis, and G. Dreyfuss. Characterization of the major hnRNP proteins from *Drosophila melanogaster. J. Cell. Biol.*, 116:257–269, 1992b.

[165] M.J. Matunis, E.L. Matunis, and G. Dreyfuss. Isolation of hnRNP complexes from *Drosophila melanogaster. J. Cell. Biol.*, 116:245–255, 1992a.

[166] M.J. Matunis, W.M. Michael, and G. Dreyfuss. Characterization and primary structure of the poly(C)-binding heterogeneous nuclear ribonucleoprotein complex K protein. *Mol. Cell. Biol.*, 12:164–171, 1992c.

[167] B. McClintock. The origin of behavior of mutable loci in maize. *PNAS*, 36:344–355, 1950.

[168] M.W. McDonnell, M.N. Simon, and F.W. Studier. Analysis of restriction fragments of T7 DNA and determination of molecular weights by electrophoresis in neutral and alkaline gels. *J. Mol. Biol.*, 110:119–146, 1977.

[169] D.A. Melton, P.A. Krieg, M.R. Rebagliati, T. Maniatis, K. Zinn, and M.R. Green. Efficient *in vitro* synthesis of biologically active RNA and RNA hybridisation probes from plasmids containing a bacteriophage SP6 promoter. *Nucleic Acids Research*, 12:7035–7056, 1984.

[170] S. Misra and D.C. Rio. Cytotype control of *Drosophila* P element tranposition: The 66kd protein is a repressor of transposase activity. *Cell*, 62:269–284, 1990.

[171] J.A. Mitchell. Fitness effects of EMS-induced mutations on the X chromosome of *Drosophila melanogaster*. I. Viability effects and heterozygous fitness effects. *Genetics*, 87:763–774, 1977.

[172] M. Mlodzik, N.E. Baker, and G.M. Rubin. Isolation and expression of *scabrous*, a gene regulation neurogenesis in *Drosophila*. *Genes Dev.*, 4:1848–1861, 1990.

[173] G. Morata and P.A. Lawrence. Control of compartment development by the *engrailed* gene in *Drosophila*. *Nature*, 255:614–617, 1975.

[174] E. Mortenson and G. Dreyfuss. RNP in maize protein. *Nature*, 337, 1989.

[175] S.M. Mount. A catalogue of splice junctions. *Nucleic Acids Research*, 10:459–471, 1982.

[176] T. Mukai and C.C. Cockerham. Spontaneous mutation rates at enzyme loci in *Drosophila melanogaster*. *PNAS*, 74:2514–2417, 1977.

[177] M.P. Mullen, C.W.J. Smith, J.G. Patton, and B. Nadal-Ginard. α-tropomyosin mutually exclusive exon selection: competition between branchpoint/polypyrimidine tracts determines default exon choice. *Genes Dev.*, 5:642–655, 1991.

[178] S.G. Nadler, J. L. Kapouch, J.I. Elliott, and K.R. Williams. Shuffling of amino acid sequence: an important control in synthetic peptide studies of nucleic acid-binding domains. *J. Biol. Chem.*, 267:3750–3757, 1992.

[179] S.G. Nadler, B.M. Merrill, W.J. Roberts, K. M. Keating, M.J. Lisbin, S.F. Barnettand S.H. Wilson, and K.R. Williams. Interactions of the A1 heterogeneous nuclear ribonucleoprotein and its proteolytic derivative, UP1, with RNA and DNA: Evidence for multiple RNA binding domains and salt-dependent binding mode transitions. *Biochem.*, 30:2968–2976, 1991.

[180] K. Nagai, C. Oubrigde, T.-H. Jessen J.Li, and P.R. Evans. Crystal structure of the RNA-binding domain of the U1 small nuclear ribonucleoprotein A. *Nature*, 348:515–520, 1990.

[181] S.-K. Oh, M.P. Scott, and P. Sarnow. Homeotic gene *Antennapedia* mRNA contains 5'-noncoding sequences that confer translational initiation by internal ribosome binding. *Genes Dev.*, 6:1643–1653, 1992.

[182] K. O'Hare, A. Driver, S. McGrath, and D.M. Johnson-Schlitz. Distribution and structure of cloned P elements from the *Drosophila melanogaster* P strain π^2. *Genet. Res. Camb.*, 60:33–41, 1992.

[183] K. O'Hare and G.M. Rubin. Structure of P transposable elements of *Drosophila melanogaster* and their sites of insertion and excision. *Cell*, 34:25–35, 1983.

[184] C.J. O'Kane and W.J. Gehring. Detection *in situ* of genomic regulatory elements in *Drosophila*. *PNAS*, 84:9123–9127, 1987.

[185] T. Özcelik, S. Leff, W. Robinson, T. Donlon, M. Lalande, E. Sanjines, A. Schinzel, and U. Francke. Small nuclear ribonucleoprotein polypeptide N (*SNRPN*), an expressed gene in the Prader-Willi syndrome critical region. *Nature Genet.*, 2:265–269, 1992.

[186] A.D. Paterson, E.S. Lander, J.D. Hewitt, S. Peterson nad S.E. Lincoln, and S.D. Tanksley. Resolution of quantitative traits into mendelian factors by using a complete linkage map of restriction fragment polymorphisms. *Nature*, 335:721–726, 1988.

[187] A.H. Paterson, J.W. DeVerns, B. Lanini, and S.D. Tanksley. Fine mapping of quantitative trait loci using selected overlapping recombinant chromosomes in an interspecies cross of tomato. *Genetics*, 124:735–742, 1990.

[188] T. Paterson, J.D. Beggs, D. Finnegan, and R. Lührmann. Polypeptide components of *Drosophila* small nuclear ribonucleoprotein particles. *Nucleic acids research*, 19:5877–5882, 1991.

[189] J.G. Patton, S.A. Mayer, P. Tempst, and B. Nadal-Ginard. Characterization and molecular cloning of polypyrimidine tract-binding protein: a component of a complex necessary for pre-mRNA splicing. *Genes Dev.*, 5:1237–1251, 1991.

[190] J.G. Patton, E.B. Porro, J. Galceran, P. Tempst, and B. Nadal-Ginard. Cloning and characterization of PSF, an novel pre-mRNA splicing factor. *Genes Dev.*, 7:393–406, 1993.

[191] W.R. Pearson and D.J. Lipman. Improved tools for biological sequence comparison. *PNAS*, 85:2444–2448, 1988.

[192] T. Pederson. Proteins associated with heterogeneous nuclear RNA in eukaryotic cells. *J. Mol. Biol.*, 83:163–183, 1974.

[193] J. Pelletier and N. Sonenberg. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poloivirus RNA. *Nature*, 334:320–325, 1988.

[194] S. Piñol-Roma, Y.D. Choi, M.J. Matunis, and G. Dreyfuss. Immunopurification of heterogeneous nuclear ribonucleoprotein particles reveals an assortment of RNA-binding proteins. *Genes Dev.*, 2:215–227, 1988.

[195] S. Piñol-Roma and G. Dreyfuss. Transcription-dependent and transcription-independent nuclear transport of hnRNP proteins. *Science*, 253:312–314, 1991.

[196] S. Piñol-Roma and G. Dreyfuss. Shuttling of pre-mRNA binding proteins between nucleus and cytoplasm. *Nature*, 355:730–732, 1992.

[197] S. Piñol-Roma, M.S. Swanson, J.G. Gall, and G. Dreyfuss. A novel heterogeneous nuclear RNP protein with a unique distribution on nascent transcripts. *J. Cell. Biol.*, 109:2575–2587, 1989.

[198] P.M. Pignatelli and T.F.C. Mackay. Hybrid dysgenesis-induced response to selection in *Drosophila melanogaster. Genet. Res. Camb.*, 54:183–195, 1989.

[199] L.R. Piper and A.E. Shrimpton. *The quantitative effects of genes which influence metric traits, in.* CAB International, Wallingford, 1989.

[200] G.J. Podgorski, J. Franke, M. Faure, and R.H. Kessin. The cyclic nucleotide phosphodiesterase gene of *Dictyostelium discoideum* utilizes alternate promoters and splicing for the synthesis of multiple mRNAs. *Mol.Cell.Biol.*, 9:3938–3950, 1989.

[201] N.J. Proudfoot and G.G. Brownlee. 3' Non-coding region sequences in eukaryotic messenger RNA. *Nature*, 263:211–214, 1976.

[202] Y. Qiu, C.-N. Chen, T. Malone, L. Richter, S.K. Beckendorf, and R.L. Davis. Characterization of the memory gene dunce of *Drosophila melanogaster. J. Mol. Biol.*, 222:553–565, 1991.

[203] K.E. Rasmusson, J.D. Raymond, and M.J. Simmons. Repression of hybrid dysgenesis in *Drosophila melanogaster* by individual naturally occurring p elements. *Genetics*, 133:605–622, 1993.

[204] G. Raychaudhuri, S.R. Haynes, and A.L. Beyer. Heterogeneous nuclear ribonucleoprotein complexes and proteins in *Drosophila melanogaster. Mol. Cell. Biol.*, 12:847–855, 1992.

[205] E.C.R. Reeve and F.W. Robertson. Studies in quantitative inheritance. *Z. indukt. Abstamm. Vererbungslehre*, 86:269–288, 1954.

[206] P.W.J. Rigby, M. Dieckmann, C. Rhodes, and P. Berg. Labeling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase I. *J. Mol. Biol.*, 113:237–251, 1977.

[207] D.C. Rio. Regulation of *Drosophila* P element transposition. *TIG*, 7:282–287, 1991.

[208] D.C. Rio, F.A.Laski, and G.M. Rubin. Identification and immunochemical analysis of biologically active *Drosophila* P element transposase. *Cell*, 44:21–32, 1986.

[209] F. Ritossa. *The bobbed locus, in The genetics and biology of* Drosophila, *ed. by M. Ashburner and E. Novitski*, volume 1b. Academic Press Inc., 1976.

[210] A. Robertson. A theory of limits in artificial selection. *Proc. Roy. Soc. London B*, 153:234–249, 1960.

[211] H.M. Robertson, C.R. Preston, R.W. Phillis, D.M. Johnson-Schlitz, W.K. Benz, and W.R. Engels. A stable source of P element transposase in *Drosophila melanogaster*. *Genetics*, 118:461–470, 1988.

[212] S. Robinow, A.R. Campos, K.-M. Yao, and K. White. The *elav* gene product of *Drosophila*, required in neurons, has three RNP consensus motifs. *Science*, 242:1570–1572, 1988.

[213] H. Roiha, G.M. Rubin, and K. O'Hare. P element insertions and rearrangements at the *singed* locus of *Drosophila melanogaster*. *Genetics*, 119:75–83, 1988.

[214] C.R. Roseland and H.A. Schneiderman. Regulation and metamorphosis of the abdominal histoblasts of *Drosophila melanogaster*. *Wilhelm Roux's Archives*, 186:235–265, 1979.

[215] G.M. Rubin, M.G. Kidwell, and P.M. Bingham. The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. *Cell*, 29:987–994, 1982.

[216] M. Russel, S. Kidd, and M.R. Kelley. An improved filamentous helper phage for generating single-stranded plasmid DNA. *Gene*, 45:333–338, 1986.

[217] A.B. Sachs, M.W. Bond, and R.D. Kornberg. A single gene from yeast for both nuclear and cytoplasmic polyadenylate-binding proteins - domain-structure and expression. *Cell*, 45:827–835, 1986.

[218] O.P. Samarina, E.M. Lukanidin, J. Molnar, and G.P. Georgiev. Structural organization of nuclear complex containing DNA-like RNA. *J. Mol. Biol.*, 33:251–263, 1968.

[219] F. Sanger, S. Nicklen, and A.R. Coulson. DNA sequencing with chain terminating inhibitors. *PNAS*, 74:5463–5467, 1977.

[220] P. Santamaria and A. Garcia-Bellido. Localization and growth pattern of the tergite anlage of *Drosophila. J. Embryol. exp. Morph.*, 28:397–417, 1972.

[221] E. Santiago, J. Albornoz, A. Domìngo, M.A. Toro, and C. Lòpez-Fanjul. The distribution of spontaneous mutations on quantitative traits and fitness in *Drosophila melanogaster. Genetics*, 132:771–781, 1992.

[222] K. Sax. The association of six differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris. Genetics*, 8:552–560, 1923.

[223] J. Schatzle, J. Bush, and J. Cardelli. Molecular cloning and characterization of the structural gene coding for the developmentally regulated lysosomal enzyme, alpha-mannosidase, in *Dictyostelium discoideum. J.Biol.Chem.*, 267:4000–4007, 1992.

[224] D. Scherly, W. Boelens, W.J. van Venrooij, N.A. Dathan, J. Hamm, and I.W. Mattaj. Identification of the RNA binding segment of human U1 A protein and definition of its binding site on U1 snRNA. *EMBO J.*, 8:4163–4170, 1989.

[225] R.E. Scossiroli and S. Scossiroli. On the relative role of mutation and recombination in responses to selection for polygenic traits in irradiated populations of *Drosophila melanogaster. J. Radiation Biol.*, 1:61–69, 1959.

[226] S.D.Tanksley, M.W. Ganal, J.P. Prince, M.C. de Vincente, M.W. Bonierbale, P. Broun, T.M. Fulton, J.J. Giovannoni, S. Grandillo, G.B.Martin, R. Messeguer, J.C. Miller, L. Miller anad A.H. Paterson, O. Pineda, M.S. Röder, R.A. Wing, W.Wu, and N.D. Young. High density molecular linkage maps of the tomato and potato genomes. *Genetics*, 132:1141–1160, 1992.

[227] L.L. Searles, A.L. Greenleaf, W.E. Kemp, and R.A. Voelker. Sites of P element insertion and structures of P element deletions in the 5' region of *Drosophila melanogaster RpII215. Mol. Cell. Biol.*, 6:3312–3319, 1986.

[228] L.L. Searles, R.S. Jokerst, P.M. Bingham, R.A. Voelker, and A.L. Greenleaf. Molecular cloning of sequences from a *Drosophila* RNA polymeraseII locus by P element transposon tagging. *Cell*, 31:585–592, 1982.

[229] A.E. Shrimpton, T.F.C. Mackay, and A.J. Leigh Brown. Transposable element-induced response to artificial selection in *Drosophila melanogaster*: Molecular analysis of selection lines. *Genetics*, 125:803–811, 1990.

[230] A.E. Shrimpton and A. Robertson. The isolation of polygenic factors controlling bristle score in *Drosophila melanogaster* I. *Genetics*, 114:437–443, 1988a.

[231] A.E. Shrimpton and A. Robertson. The isolation of polygenic factors controlling bristle score in *Drosophila melanogaster* II. *Genetics*, 114:445–459, 1988b.

[232] C.W. Siebel and D.C. Rio. Regulated splicing of the *Drosophila* P transposable element third intron *in vitro*. *Science*, 248:1200–1208, 1990.

[233] P.T. Sillekens, W.J. Habets, R.P. Beijer, and W.J. Van Venrooij. cdna cloning of the human U1 snRNA-associated A protein. *EMBO J.*, 6:3841–3848, 1987.

[234] M.J. Simmons and L.M. Bucholz. Transposase titration in *Drosophila melanogaster*: A model of cytotype in the P-M system of hybrid dysgenesis. *PNAS*, 82:8119–8123, 1985.

[235] M.J. Simmons and J.K. Lim. Site specificity of mutations arising in dysgenic hybrids of *Drosophila melanogaster*. *PNAS*, 77:6042–6046, 1980.

[236] M.J. Simmons, E.W. Sheldon, and J.F. Crow. Heterozygous effects on fitness of EMS-treated chromosomes in *Drosophila melanogaster*. *Genetics*, 88:575–590, 1978.

[237] P. Simpson. Lateral inhibition and the development of the sensory bristles of the adult peripheral nervous system of *Drosophila*. *Development*, 109:509–519, 1990.

[238] S.T. Smale and D. Baltimore. The "Initiator" as a transcription control element. *Cell*, 57:103–113, 1989.

[239] J. Maynard Smith and K.C. Sondhi. The arrangement of bristles in *Drosophila*. *J. Embryol. exp. Morph.*, 9:661–672, 1961.

[240] D. Smoller, C. Friedel, A. Schmid, D. Bettler, L.Lam, and B. Yedvobnick. The *Drosophila* neurogenic locus *mastermind* encodes a nuclear protein unusually rich in amino acid homopolymers. *Genes Dev.*, 4:1688–1700, 1990.

[241] K.C. Sondhi. Genetic control of an anteroposterior gradient and its bearing on structural orientation in Drosophila. *Genetics*, 51:653–657, 1964.

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

[243] S.G. Spickett. Genetic and developmental studies of a quantitative character. *Nature*, 199:870–873, 1963.

[244] S.G. Spickett and J.M. Thoday. Regular responses to selection. *Genet. Res.*, 7:96–121, 1966.

[245] A.C. Spradling and G.M. Rubin. *Drosophila* genome organisation: conserved and dynamic aspects. *Ann. Rev. Genet.*, 15:219–264, 1981.

[246] A.C. Spradling and G.M. Rubin. Transposition of cloned P elements into *Drosophila* germ line chromosomes. *Science*, 218:341–347, 1992.

[247] A.H. Sturtevant. Studies on the bristle pattern of *Drosophila*. *Developmental Biol.*, 21:48–61, 1970.

[248] J.A. Sved. Hybrid dysgenesis in *Drosophila melanogaster*: a possible explanation in terms of spatial organisation of chromosomes. *Aust. J. Biol. Sci.*, 29:375–388, 1976.

[249] M.S. Swanson and G. Dreyfuss. Classification and purification of proteins of heterogeneus nuclear ribonucleoprotein particles of RNA-binding specificities. *Mol. Cell. Biol.*, 8:2237–2241, 1988a.

[250] M.S. Swanson and G. Dreyfuss. RNA binding specificity of hnRNP proteins: a subset bind to the 3' end of introns. *EMBO J.*, 11:3519–3529, 1988b.

[251] M.S. Swanson, T.Y. Nakagawa, K. LeVan, and G. Dreyfuss. Primary structure of human nuclear ribonucleoprotein particle C proteins: Conservation of sequence and domain structures in heterogeneous nuclear RNA, mRNA and pre-rRNA-binding proteins. *Mol. Cell. Biol.*, 7:1731–1739, 1987.

[252] A. Szabo, J. Dalmau, G.Manley, M. Rosenfeld, E. Wong, J. Henson, J.B. Posner, and H. M. Furneaux. Hud, a paraneoplastic encephalomyelitis antigen, contains RNA-binding domains and is homologous to elav and sex-lethal. *Cell*, 67:325–333, 1991.

[253] W. E. Theurkauf, H. Baum, J. Bo, and P.C. Wensink. Tissue-specific and constitutive α-tubulin genes of *Drosophila melanogaster* code for structurally distinct proteins. *PNAS*, 83:8477–8481, 1986.

[254] J.M. Thoday. Location of polygenes. *Nature*, 191:368–370, 1961.

[255] J.M. Thoday and J.N. Thompson. The number of segregating genes implied by contiuous variation. *Genetica*, 46:335–344, 1976.

[256] J.O. Thomas, K. Glowacka, and W. Szer. Structure of complexes between a major protein of heterogeneous nuclear ribonucleoprotein particles and polyribonucleotides. *Mol. Biol.*, 171:439–455, 1983.

[257] J.N. Thompson. Quantitative variation and gene number. *Nature*, 258:665–668, 1975.

[258] A. Torkamanzehi, C. Moran, and F.W. Nicholas. P-element-induced mutation and quantitative variation in *Drosophila melanogaster*: lack of enhanced response to selelction in lines derived from dysgenic crosses. *Genet. Res.*, 51:231–238, 1988.

[259] A. Torkamanzehi, C. Moran, and F.W. Nicholas. P element transposition contributes substantial new variation for a quantitative trait in *Drosophila melanogaster*. *Genetics*, 131:73–78, 1992.

[260] S. Tsubota, M. Ashburner, and P. Schedl. P-element-induced control mutations at the *r* gene of *Drosophila melanogaster*. *Mol. Cell. Biol.*, 5:2567–2574, 1985.

[261] S. Tsubota and P.Schedl. Hybrid dysgenesis-induced revertants of insertions at the 5' end of *rudimentary* gene in D*rosophila melanogaster*: transposon-induced control mutations. *Genetics*, 114:165–182, 1986.

[262] R.A. Voelker, W. Gibson, J.P. Graves, J.F. Sterling, and M.T. Eisenberg. The *Drosophila suppressor of sable* gene encodes a polypeptide with regions similar to those of RNA-binding proteins. *Mol. Cell. Biol.*, 11:894–905, 1991.

[263] J. Wang and T. Pederson. A 62,000 molecular weight spliceosome protein crosslinks to the intron polypyrimidine tract. *Nucleic Acids Research*, 18:5995–6001, 1990.

[264] E. Wieschaus and C. Nüsslein-Volhard. *Looking at embryos, in Drosophila, a practical approach, ed. D.B.Roberts.* IRL Press, Oxford, 1986.

[265] P.R. Winship. An improved method for directly sequencing PCR amplified material using dimethyl sulphoxide. *Nucleic Acids Research*, 17:1266, 1989.

[266] M. Wittekind, M. Görlach, M. Friedrichs, G. Dreyfuss, and L. Mueller. 1H, 13C, and 15N NMR assignments and global folding pattern of the RNA-binding domain of the human hnRNP C proteins. *Biochem.*, 31:6254–6265, 1992.

[267] S. Wright. *The Genetics of Quantitative Variability, in Quantitative Inheritance, eds. E.C.R. Reeve and C.H. Waddington.* Her Majesty's Stationary Office, London, 1952.

[268] X. Yang, K.T. Seow, S.M. Bahri, S.H. Oon, and W. Chia. Two Drosophila receptor-like tyrosine phosphatase genes are expressed in a subset of developing axons and pioneer neurons in the embryonic CNS. *Cell*, 67:661–673, 1991.

[269] K.-M. Yao and K. White. Organizational analysis of *elav* gene and functional analysis of elav protein of *Drosophila melanogaster* and *Drosophila virilis. Mol. Cell. Biol.*, 11:2994–3000, 1991.

[270] B.H. Yoo. Long-term selection for a quantitative character in large replicate populations of *Drosophila melanogaster. Genet. Res.*, 35:19–31, 1980.

[271] Z. Zachar, T.-B. Chou, and P.M. Bingham. Evidence that a regulatory gene autoregulates splicing of its transcript. *EMBO J.*, 6:4105–4111, 1987.

[272] A.M. Zahler, W.S. Lane, J.A. Stolk, and M.B. Roth. SR proteins: a conserved family of pre-mRNA splicing factors. *Genes Dev.*, 6:837–847, 1992.

[273] P.D. Zamore, J.G. Patton, and M.R. Green. Cloning and domain structure of the mammalian splicing factor U2AF. *Nature*, 355:609–614, 1992.

[274] Z.-B. Zeng. Correcting the bias the WRIGHT's estimates of the number of genes affecting a quantitative character: A further improved method. *Genetics*, 131:987–1001, 1992.