CHARACTERIZATION OF THE GENOME OF *ARABIDOPSIS THALIANA*

Thesis by

Robert Edwin Pruitt

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1986

(Submitted December 10, 1985)

There is something fascinating about science. One gets such wholesale returns of conjecture out of such a trifling investment of fact.

- Mark Twain

ACKNOWLEDGEMENTS

First I would like to thank my thesis advisor, Elliot Meyerowitz, for having the faith and understanding to let me work on *Arabidopsis* at a time when he was working to establish a thriving laboratory which still works predominantly on *Drosophila*.

Secondly, I would like to thank all the members of the Meyerowitz lab who lent their time, skills and advice and are in part responsible for the existence of this work. In particular, I would like to thank Leslie Leutwiler, who made the early work so much more enjoyable and also provided the first real impetus to finding out what the *Arabidopsis* genome was really like; and Chris Martin and K. VijayRaghavan, who helped to pull it all together when I got in over my head.

Lastly, I would like to thank my wife Lynn and daughter Laura Jean who have somehow managed to coexist with me and also to take on my share of all the joint responsibilities for the last several months.

ABSTRACT

The work described in this thesis consists of two different types of characterizations of the genome of *Arabidopsis thaliana*. The first part (Chapter 2) is concerned with the organization of the genome as a whole while the second part (Chapter 3) is concerned with a detailed analysis of two pairs of genes which are expressed in the developing *Arabidopsis* seed. The analysis presented in Chapter 2 is based on a characterization of the DNA sequences found in 50 randomly selected recombinant lambda clones. The clones were characterized by various blotting and restriction digestion techniques to determine their content of unique and repetitive DNA and the interspersion pattern of these two types of sequences. The various repetitive sequences which were identified were characterized further as to the exact nature of the repeated sequence. The primary conclusion which can be drawn from this analysis is that the *Arabidopsis* genome consists predominantly of long contiguous blocks of single-copy sequences. The analysis presented in Chapter 3 is concerned with two pairs of genes which are expressed specifically and abundantly in the developing seed of *Arabidopsis*. These genes are characterized with respect to the time of their expression during seed development, the organization of the regions of the genome containing the genes and the directions the genes are transcribed. The nucleotide sequences of the genes and flanking regions are also presented.

TABLE OF CONTENTS

Chapter 1

Arabidopsis thaliana and Plant Molecular Genetics

(reprinted from *Science* **229:** 1214-1218.)

ARABIDOPSIS THALIANA AND PLANT MOLECULAR GENETICS

Elliot M. Meyerowitz and Robert E. Pruitt

SUMMARY

Arabidopsis thaliana is a small flowering plant with a variety of properties that make it an excellent organism for experiments in molecular genetics. Among these are a tiny nuclear genome, a near absence of dispersed repetitive DNA, and a four- to five-week generation time. In addition, mutations that affect hormone synthesis and response, many different enzyme activities, and numerous developmental processes have been isolated and characterized.

It is worth understanding the molecular genetics of plants not only because of the practical value of such understanding in the improvement of crops, but also because studies of plants offer an opportunity to gain insights into a variety of basic life processes that are unique to plants. Plants do not use the same hormones as animals, and do not use hormones in the same way that animals use them. Most or all of the cells in a plant may both produce and respond to plant hormones at some times. Plants respond to stress differently than animals and respond to light in diverse and subtle ways. Photosynthesis is just one of the light responses of plants that is not found in animals. Even the basic developmental processes of plants have features that distinguish them from those of animals. Among the differences are the absence of cell migration in plant development, and the fact that parts of each flowering plant (the meristems) remain embryonic and can produce adult organs, including germ cells, throughout the life of the plant. Further, individual differentiated cells taken from vegetative plant parts can dedifferentiate and regenerate to form entire, fertile plants.

While it is sensible and necessary to study crop plants for purposes of crop improvement, the crop plants now used for basic classical and molecular genetic studies have disadvantages for some of the types of experimentation used in this work. Classical genetics depends on the ability to raise many successive generations of organisms in large numbers. Typical crop plants have generation times of several months, and require a great deal of field space for growth in large numbers. The genetics of some of these plants is also made more difficult by polyploidy or allopolyploidy. The ease with which recombinant DNA work can be done with any organism depends, in part, on the size of its nuclear genome: the smaller the genome, the less work required to screen recombinant DNA libraries and thus to isolate any particular gene. The genome sizes of the plants presently

used for recombinant DNA work are large, and similar to that found in mammals (Table 1). In addition, the plants generally used for recombinant DNA work contain large amounts of dispersed repetitive DNA (1), which makes such procedures as genomic blot analysis with genomic clones and chromosome walking extremely difficult, or impossible. When starting plant work, we resolved to find an organism that would have as few of these experimental disadvantages as possible. It is important to state that a great deal of excellent work in classical and molecular genetics has been done and is now being done using traditional crop plants; still, we felt that a system without the specific disadvantages outlined above would offer the possibility for doing certain types of experiments more rapidly, more easily, and at less expense than current systems permit. The plant we have chosen is *Arabidopsis thaliana*.

CLASSICAL AND BIOCHEMICAL GENETICS

Arabidopsis is a plant that has been used in classical genetic work for over 40 years (2), and on which there is an extensive genetic and ecological literature. It is a member of the mustard family, along with some more familiar plants such as cabbages and radishes. *Arabidopsis* is a harmless weed of no food or economic value; nonetheless, the literature on the plant indicates that it might be of considerable value in basic research. First, it is well suited to classical genetic work: it has a generation time of only 5 weeks; each individual plant can produce more than 10,000 seeds; and it is of such small size that dozens of plants can be grown in a small pot, and tens of thousands in a small room (3). It requires no more than moist soil and a fluorescent light for rapid growth. The small flowers contain both anthers and pistils, and the plant typically self- fertilizes. This allows new mutations to be made homozygous with minimal effort. When desired,

cross-fertilization can be simply and rapidly effected as well, allowing the crosses necessary for genetic mapping and for the production of multiple mutant stocks. Mutagenesis can be performed by soaking seeds in chemical mutagens such as ethyl methanesulfonate (EMS) or by irradiation of seeds soaked in water. The mutagenized seeds are planted, grown to maturity and allowed to self-fertilize, resulting in seeds which are homozygous for new mutations.

A variety of mutations have been isolated in many different laboratories. Visible mutations useful as markers in genetic mapping have been isolated, with phenotypes affecting every part of the plant. These include those affecting the wax coat of the epidermal cells (*cer* mutants, 4) and the trichomes normally found on leaves and stems (*dis* and *gl* mutants, 5) as well as those causing more easily visible effects such as changes in flower morphology (*ag*, *ap* and *pi* mutants, 6) and growth habit (*er* and *cp* mutants, 7). Other mutations affect the embryonic development of *Arabidopsis*. These lead to embryonic lethality, with the developmental stage and phenotype of embryonic developmental arrest depending on the specific gene mutated (8). Many leaf color mutants exist, including two that give variegated plants. Plants which are homozygous for the immutans (*im*) mutation have leaves which are variegated white and green; the degree of this variegation depends both on the particular allele of the *im* mutation and on the intensity of light falling on the plant at the time a given leaf develops. This recessive mutation is inherited from both white and green sectors (9). Plants which are homozygous for the chloroplast mutator (*chm*) mutation also have leaves which bear white and green sectors. In this case the white sectors are due to failure of the chloroplasts to develop normally. Although plants which are heterozygous for the *chm* mutation do not produce aberrant chloroplasts, chloroplasts which are maternally inherited from homozygous *chm* plants by heterozygous plants continue to be of abnormal structure. In fact, it is possible to

isolate from these out-crossed *chm* strains stable homoplastidic lines of *Arabidopsis* in which all of the chloroplasts have the same abnormal appearance. It has been proposed that the nuclear *chm* mutation acts by causing errors during chloroplast DNA replication (10).

Biochemical mutants of various types have also been isolated. Mutations in several genes give rise to plants requiring thiamine or the thiamine precursors 2,5-dimethyl-4- aminopyrimidine or 4-methyl-5-hydroxyethyl thiazole for growth (11,12). Somerville and Ogren have characterized biochemical mutations that affect photorespiration and photosynthesis. These include nuclear mutations which result in the absence of activity of chloroplast glutamate synthase, phosphoglycolate phosphatase, mitochondrial glycine decarboxylase, mitochondrial serine transhydroxymethylase and serine-glyoxylate aminotransferase (13). Mutations which affect the activity of nitrate reductase, and of alcohol dehydrogenase, have also been characterized (14).

Mutations have also been obtained which appear to disrupt the normal action of several of the plant growth regulators. Koornneef and his collaborators have recovered mutations which result in the absence of gibberellins or of abscisic acid as well as mutations conferring resistance to high levels of exogenous abscisic acid. The mutations which affect gibberellins fall into five complementation groups. Three of these contain alleles which, when homozygous, prevent germination unless the seeds are treated with exogenous gibberellins (15). A mutant strain which lacks abscisic acid has been used by Karssen et al. (16) to examine the relative effects of maternal and embryo-derived abscisic acid on seed dormancy. Mutations which result in resistance to high levels of a synthetic auxin have also been recovered, and some alleles of these mutations produce plants which have agravitropic roots (17). One of these mutations (*Dwf*) is dominant and produces a plant with a dwarf rosette and inflorescence when

heterozygous, and is lethal when homozygous.

Over 75 of the known mutations have been assembled into a genetic linkage map by Koornneef et al. (6). This genetic map consists of five linkage groups, in concordance with the haploid chromosome number of five. The location of the centromeres on the genetic map has been determined by the use of strains which are trisomic for one arm of a chromosome (6, 18). *Arabidopsis* stocks containing many of the mutations on the genetic map in combinations suitable for mapping specific chromosome regions, and many strains collected from the wild, as well, are available from a central international seed collection (19).

In addition to growth in soil, it is also possible to grow whole *Arabidopsis* plants on sterile, biochemically defined media; both solid and liquid media have been used (11, 20). *Arabidopsis* cells have also been grown in tissue culture, and it has been possible to regenerate plants from such cells (21). With tissue culture it has been possible to isolate some types of biochemical mutations by selection; it may also be possible to use biochemical selections to assay transformation of tissue culture cells.

MOLECULAR GENETICS: GENOME SIZE AND ORGANIZATION

When our work began, two lines of evidence indicated that *Arabidopsis* had a small genome. Microspectrophotometry of Feulgen stained nuclei indicated that the quantity of DNA in the haploid genome of *Arabidopsis* was 0.2 pg, or approximately 2 $x10^8$ base pairs, making it the smallest angiosperm genome which had been measured at that time (22). Another line of evidence indicating that *Arabidopsis* might have a small genome is based on the proportionality of genome size and nuclear volume. This relationship implied that *Arabidopsis* should have a haploid genome size of roughly 10^9 base pairs (23). These estimates indicate that

the *Arabidopsis* genome is small enough to greatly simplify the task of screening recombinant DNA libraries to isolate genes.

We have performed a more detailed analysis of the *Arabidopsis* genome to verify its small size and to determine the fraction and nature of repetitive sequences present. Our initial study was a reassociation analysis of DNA extracted from whole plants (24). This gave an estimated genome size of 7×10^7 base pairs, a size only five times that of the yeast genome (25) and much smaller than that of other flowering plants (Table 1). This analysis also showed that the whole plant DNA contained 10-14% very rapidly reannealing sequences (either highly repeated sequences or inverted repeat sequences) and 23-27% middle repetitive sequences. By using a labeled tracer containing cloned chloroplast sequences it was demonstrated that almost all of the middle repetitive sequences were from the chloroplast genome, and thus that *Arabidopsis* has far less nuclear repetitive DNA than other angiosperms (1, Table 2). Leutwiler et al. (24) also examined the degree of cytosine methylation in *Arabidopsis* and found that only 4.6% of the cytosines derived from whole plant DNA were present as 5-methyl cytosine, the lowest value known for a flowering plant.

More recently we have used recombinant DNA techniques to examine random segments of genomic DNA (26). Randomly selected recombinant lambda phage were subjected to a variety of experimental manipulations to determine if they contain unique or repetitive sequences (or both), and to characterize the nature of the repetitive sequences that were contained in the clones. These experiments confirm and extend our previous analysis. A majority of the 50 clones analyzed contain only unique sequences and most of the clones which contain repetitive sequences are derived either from the chloroplast genome or from the nuclear DNA that codes for the large ribosomal RNAs. There are approximately 570 copies of the ribosomal DNA per haploid genome, each

ribosomal DNA repeat unit is about 10 kb in length, and the repeats are largely arranged in tandem array. In addition there are two clones which appear to contain duplicated sequences, one clone which contains a low copy number repeat unit that is apparently conserved in all copies (this may represent a fragment of mitochondrial DNA), and three clones which contain both repeated sequences and unique sequences interspersed with each other. These results indicate that the *Arabidopsis* nuclear genome consists of predominantly unique sequences and that most of the nuclear repetitive DNA is ribosomal DNA. The experiments clearly indicate that much of the nuclear DNA of *Arabidopsis* is organized as extremely long blocks (on average 120 kb) of unique sequences.

These two sets of experiments indicate that *Arabidopsis* has a remarkably small and simple genome. This fact, taken together with the results from all of the previous work on *Arabidopsis* indicates that this plant does represent a very good model system for basic research in plant molecular biology.

MOLECULAR GENETICS: SPECIFIC GENES

We have started to do additional work studying individual genes from *Arabidopsis* to provide material for analysis of the mechanisms by which gene expression is regulated by both developmental and environmental stimuli. So far we have examined three different genes or gene families and they have led us to two generalizations that emphasize the utility of *Arabidopsis* for experiments in molecular cloning. The first is that it is possible to cross-hybridize genes from a wide variety of flowering plants with the homologous gene or genes in *Arabidopsis*. This is perhaps not surprising, since fossil evidence indicates that the first major radiation of the angiosperms took place between 1 and 1.5 x 10^8 years ago, when the mammals, too, were first radiating. The second generalization that

our experience so far allows is that proteins which are encoded by multiple genes or large gene families in other plants are encoded by single genes or small gene families in *Arabidopsis*.

The first gene we have examined is that encoding the large seed storage protein (27). We have cloned this gene from a recombinant library containing genomic DNA by using a cDNA clone encoding the 12S seed storage protein of *Brassica napus* (28). In the Columbia strain of *Arabidopsis* this protein is encoded by a single gene indicating that the heterogeneity observed in other storage protein gene families is not required in this plant. Further, the fact that there is only one gene coding for this protein makes it clear that the tissue and time-specific regulation of the protein is due to the activity of only one gene, and that experiments to understand this regulated gene expression need only deal with this single sequence, and not a large family of similar genes, as in other plants.

A second series of experiments (29) involved the cloning of the genes encoding the light-induced chlorophyll a/b binding protein of *Arabidopsis*. Chlorophyll a/b binding protein, or light-harvesting chlorophyll protein, is a nuclear-coded component of the light-harvesting antenna complex of the chloroplast thylakoid membranes. These genes were isolated from an *Arabidopsis* genomic library using as a probe a genomic clone from the tiny aquatic monocot *Lemna gibba* (30). There are three genes encoding this protein in *Arabidopsis*, all located in a small gene cluster of less than 6,000 base pairs. The three genes of the *Arabidopsis* chlorophyll a/b binding protein family contrast with the much larger number in the homologous families in other plants. An example is Petunia, where the thorough studies of Dunsmuir et al. (31) have shown that there may be as many as 16 or more members of this gene family, divided into at least 5 major subclasses. Perhaps even more surprisingly, it has been determined by DNA sequencing that after the transit peptides of the three *Arabidopsis* proteins are

removed all of the genes encode an identical product. Thus, as with the seed storage protein, there does not appear to be any protein heterogeneity for the chlorophyll a/b binding protein encoded in the genome of *Arabidopsis*. Any model for this gene family in *Arabidopsis* must recognize that the purpose of the multiple genes is not to provide a series of variant proteins of slightly different function. An (32) has taken advantage of the small number of *Arabidopsis* chlorophyll a/b binding protein genes: since there are only three copies, it is possible in a short series of experiments to test all of the upstream regulatory regions of the genes for function and light inducibility. He has so far fused two of these regions to bacterial chloramphenicol acetyltransferase genes and used *Agrobacterium* Ti-plasmid constructs to introduce the fusion genes to tissue culture cells of tobacco. In some instances each of the two *Arabidopsis* DNA fragments have conferred light-inducibility on the bacterial gene.

One final example of a specific *Arabidopsis* gene that has been cloned and characterized is that coding for alcohol dehydrogenase (33). A classical genetic study of alcohol dehydrogenase allozymes had already indicated that there was only a single ADH gene in *Arabidopsis*, in contrast to the two or three in many other plants (34). The *Arabidopsis* gene was isolated from a recombinant library by cross-hybridization with a labeled maize ADH1 gene fragment. The gene is single-copy, in confirmation of the earlier genetic results; DNA sequencing shows that the *Arabidopsis* gene shares more than 70% nucleic acid homology with the maize ADH1 gene. The proteins coded by the *Arabidopsis* gene and the maize ADH1 gene have slightly more than 80% amino acid identity. Further, the structure of the *Arabidopsis* and maize genes are strikingly similar. Both the maize ADH1 and ADH2 genes contain nine intervening sequences at identical positions (35). The *Arabidopsis* gene has six intervening sequences; the positions of all six are coincident with the corresponding positions of six of the maize

introns.

Thus, the existing evidence supports the conclusions that *Arabidopsis* genes will cross-hybridize with homologous genes from other angiosperms, both monocots and dicots, and that genes found in large gene families in other plants can exist in small families, or in single copy, in *Arabidopsis*. The practical importance of the first conclusion is that genes of interest can be simply cloned from the extraordinarily small *Arabidopsis* genome, then used as probes for the isolation of the homologous genes from plants of economic value. The importance of the second conclusion is that genes found in gene families can be more easily and more thoroughly studied in *Arabidopsis* than in other angiosperms.

THE FUTURE OF *ARABIDOPSIS* RESEARCH

For *Arabidopsis* to be truly useful as a tool for molecular genetic research, two additional techniques must be developed. It must be possible to clone genes about which no more is known than their mutant phenotype, and it must be possible to introduce cloned genes which have been modified *in vitro* back into the plant, in order to assay their *in vivo* function. The second of these techniques is currently being pursued by many laboratories using various techniques. *Arabidopsis* is known to be susceptible to infection by *Agrobacterium tumefaciens*, and it is known that Ti-plasmid strains of *Agrobacterium* cause typical tumors on *Arabidopsis* (36). It will very likely prove possible to introduce cloned sequences into the *Arabidopsis* genome using the same methods currently used in the transformation of other plants. If a method can be found which works efficiently enough, it may be possible to take advantage of the small genome of *Arabidopsis* in isolation of genes by "shotgun" transformation with a genomic library cloned in a Ti plasmid-based vector. This would involve transformation of mutant plants or

plant cells with a complete random recombinant library derived from wild type plant DNA, then assaying individual transformed cells or plants for complementation of the mutant phenotype by the expression of the introduced DNA. Using a cosmid vector with a capacity of 20- 25 kb it would require approximately 3,000 transformation events to introduce one genome equivalent of wild type DNA to a set of mutant plants or cells, and 14,000 to introduce the 4.6 genomes necessary to have a 99% chance of recovering any specific gene. Even with an efficient technique these are large numbers of transformation events unless the gene of interest has a phenotype which can be selected directly.

An alternative approach is to use a procedure of successive isolations of overlapping cloned segments starting from a known cloned genetic location and proceeding to any nearby genetic locus. The *Arabidopsis* genome is unique among the flowering plants in its suitability for this sort of process, due to its small size, and particularly due to the near absence of dispersed repeated sequences. Each of these sequences would be a fork in the path from the starting clone to the eventual goal; no other flowering plant is known to have a genome with few enough of these sequences, spaced far enough from each other, to allow such a procedure to be practical (Table 2). In order for a successive isolation procedure to be used, it is necessary to have starting points which are known to be located relatively close to the gene of interest. In order to provide these starting points we are currently involved in the production of a genetic map using restriction fragment length polymorphisms (RFLPs, 37). It is easy to find RFLPs in different wild-type strains of *Arabidopsis*. We have tested a number of different strains and selected the Niederzenz strain (originally collected in Niederzenz, West Germany) and the Landsberg strain (initially collected in Landsberg, East Germany) as the parental strains. Landsberg was chosen because most of the mutations used in construction of the published genetic maps of *Arabidopsis* are in the Landsberg

genetic background. Niederzenz was selected as the second parental strain because it grows as quickly as Landsberg (many other wild strains do not) and because it shows many RFLPs when compared to Landsberg. We have crossed these two parental strains, selfed the F1, and collected seeds from self-fertilized F2 plants. These pools of F3 seeds have been grown up and DNA prepared from them, representing the alleles present in the F2 plant from which they descended. By genome blot analysis it will be possible to determine the genotype of each of the F2 plants at each polymorphic locus. From these data, linkage distances can be calculated just as for any other pair of genetic markers. By including visible markers in the original cross and performing subsequent crosses with other visible markers it will be possible to align the RFLP map with the published genetic map of visible mutations. The clones used as probes on the RFLP genome blots can then serve as starting points for successive clone isolations.

In order to tell when the desired gene has been cloned, it will be necessary to transform cloned DNA segments into the plant and to assay them for the ability to complement mutations in the gene. Since only a small number of clones close to the gene of interest need to be tested by introduction into the plant genome, this procedure will not require the high frequency of transformation demanded by the first general gene isolation method described.

The ability to transform *Arabidopsis* not only has the potential for allowing cloning of any gene having a known mutant phenotype, but will also allow the type of detailed analysis of these genes that is now being performed with genes of yeast, Drosophila and mice. Among the genes that might be cloned and analyzed are those affecting particular enzyme activities, those with specific effects on the ability of *Arabidopsis* to synthesize and respond to plant hormones, and those whose mutations give specific developmental abnormalities. Our hope is that

Arabidopsis will soon join the other organisms with which a combined genetic and molecular approach has led to both fundamental and practical scientific advances.

Plant	Haploid Genome Size (kilobase pairs)	Lambda Clones in complete library
Arabidopsis	70,000	16,000
Mung bean	470,000	110,000
Cotton	780,000	180,000
Tobacco	1,600,000	370,000
Soybean	1,800,000	440,000
Pea	4,500,000	1,000,000
Wheat	5,900,000	1,400,000

Table 1. Haploid genome size in various flowering plants, and the number of lambda clones that must be screened to have a 99% chance of isolating a single-copy sequence from these genomes. The genome sizes are calculated from kinetic complexity measurements (38). The library sizes are calculated assuming a random nuclear DNA library with an average clone insert length of 20 kilobase pairs. 4.6 genome equivalents must be screened for a 99% probability of isolating any individual unique sequence.

Plant	Average Size in the Predominant Class of Single Copy Sequences (in kilobase pairs)	Amount of Repetitive DNA in Haploid Genome (in kilobase pairs)
Arabidopsis	120	18,000
Mung Bean	6.7	160,000
Cotton	1.8	310,000
Tobacco	1.4	1,200,000
Soybean	3	1,100,000
Pea	0.3	3,800,000
Wheat	1	4,400,000

Table 2. Average size in kilobase pairs of the single-copy DNA sequences interspersed with repetitive sequences, and total amount of repetitive DNA in various plant genomes. The *Arabidopsis* calculations are from (24) for amount of repetitive DNA, and (26) for size of unique DNA stretches; the unique sequence size is from measurements of random cloned fragments. For the other plants this measurement is from reassociation analysis (38).

REFERENCES AND NOTES

1. R. Flavell, Ann. Rev. *Plant Physiol.* **31**, 569-596 (1980).

2. F. Laibach, *Bot. Arch.* **44**, 439-455 (1943); _____, *Naturwiss.* **31**, 246 (1943); E. Reinholz, *Naturwiss.* **34**, 26-28 (1947).

3. G.P. Redei, *Ann. Rev. Genet.* **9**, 111-127 (1975).

4. L.M.W. Dellaert, J.Y.P. Van Es and M. Koornneef, *Arabidopsis Information Service* **16**, 10-26 (1979).

5. W.J. Feenstra, *Arabidopsis Information Service* **15**, 35-38 (1978); S. Lee-Chen and L.M. Steinitz-Sears, *Can. J. Genet. Cytol.* **9**, 381-384 (1967); M. Koornneef, L.W.M. Dellaert and J.H. van der Veen, *Mut. Res.* **93**, 109-123 (1982).

6. M. Koornneef, J. van Eden, C. J. Hanhart, P. Stam, F. J. Braaksma and W. J. Feenstra, *J. Hered.* **74**, 265-272 (1983).

7. G.P. Redei, *Z. Vererbungsl.* **93**, 164-170 (1962); M. Koornneef, J. van Eden, C. J. Hanhart, P. Stam, F. J. Braaksma and W.J. Feenstra, *op. cit.*

8. D.W. Meinke and I.M. Sussex, *Dev. Biol.* **72**, 50-61 (1979); D.W. Meinke, *Theor. Appl. Genet.* **69**, in press.

9. G.P. Redei, *Genetics* **56**, 431-443 (1967); G. Robbelen, *Planta* (Berl.) **80**, 237-254 (1968).

10. G.P. Redei, *Mut. Res.* **18**, 149-162 (1973).

11. J. Langridge, *Nature* **176**, 260-261 (1955).

12. S.L. Li and G.P. Redei, *Biochem. Genet.* **3**, 163-170 (1969).

13. C.R. Somerville and W.L. Ogren, *Nature* **286**, 257-259 (1980); ____, *ibid.* **280**, 833-836 (1979); _____, *Biochem. J.* **202**, 373-380 (1982); ___, *Plant Physiol.* **67**, 666-671 (1981); ___, *Proc. Natl. Acad. Sci. USA* **77**, 2684-2687 (1980).

14. F.J. Braaksma and W.J. Feenstra, *Theor. Appl. Genet.* **64**, 83- 90 (1982); M. Jacobs and D. Schwartz, *Arabidopsis Information Service* **17**, 88-90 (1980).

15. M. Koornneef and J.H. van der Veen, *Theor. Appl. Genet.* **58**, 257-263 (1980); M. Koornneef, M.L. Jorna, D.L.C. Brinkhorst-van der Swan and C.M. Karssen, *Theor. Appl. Genet.* **61**, 385-393 (1982); M. Koornneef, G. Reuling and C.M. Karssen, *Physiol. Plant.* **61**, 377-383 (1984).

16. C.M. Karssen, D.L.C. Brinkhorst-van der Swan, A.E. Breekland and M. Koornneef, *Planta* **157**, 158-165 (1983).

17. J.I. Mirza, G.M. Olsen, T.-H. Iversen and E.P. Maher, *Physiol. Plant.* **60**, 516-522 (1984); G.M. Olsen, J.I. Mirza, E.P. Maher and T.-H. Iversen, *ibid.* 523-531 (1984).

18. M. Koornneef and J.H. Van der Veen, *Genetica* **61**, 41-46 (1983).

19. The seed collection is maintained by Prof. Dr. A.R. Kranz, Botanisches Institut, J.W. Goethe-Universitat, Siesmayerstrasse 70, Postfach 111 932, D-6000 Frankfurt am Main 11, Bundesrepublik Deutschland. Prof. Dr. Kranz also edits and distributes *Arabidopsis Information Service*, an annual newsletter that includes original contributions, reviews, reports of new mutations and lists of available stocks.

20. G.P. Redei and C.M. Perry, *Arabidopsis Information Service* **8**, 34 (1971); N. Goto, *ibid.* 19, 55-62 (1982).

21. I. Negrutiu, M. Jacobs and W. de Greef, *Z. Pflanzenphysiol.* **90**, 363-372 (1978).

22. M.D. Bennett and J.B. Smith, *Proc. Roy. Soc. Lond.* B **274**, 227-274 (1976).

23. A.H. Sparrow, H.J. Price and A.G. Underbrink, *Brookhaven Symp. Biol.* **23**, 451-494 (1972).

24. L.S. Leutwiler, B.R. Hough-Evans and E.M. Meyerowitz, *Mol. Gen. Genet.* **194**, 15-23 (1984).

25. G.D. Lauer, T.M. Roberts and L.C. Klotz, *J. Mol. Biol.* **114**, 507-526 (1977).

26. R.E. Pruitt and E.M. Meyerowitz, *J. Mol. Biol.*, in press.

27. R.E. Pruitt and E.M. Meyerowitz, manuscript in preparation.

28. A.E. Simon, K.M. Tenbarge, S.R. Scofield, R.R. Finkelstein and M.L. Crouch, *Plant Mol. Biol.* **5**, 191-201 (1985).

29. L.S. Leutwiler, E.M. Meyerowitz and E.M. Tobin, in preparation.

30. W.J. Stiekema. C.F. Wimpee, J. Silverthorne and E.M. Tobin, *Plant Physiol.* **72**, 717-724 (1983).

31. P. Dunsmuir, S.M. Smith and J. Bedbrook, *J. Mol. Appl. Genet.* **2**, 285-300 (1983).

32. G. An, personal communication.

33. C. Chang and E.M. Meyerowitz, *Proc. Natl. Acad. Sci USA*, in press.

34. R. Dolferus and M. Jacobs, *Biochem. Genet.* **22**, 817-838 (1984).

35. E.S. Dennis et al., *Nucl. Acids Res.* **12**, 3983-4000 (1984); E.S. Dennis, M.M. Sachs, W.L. Gerlach, E.J. Finnegan and W.J. Peacock, *ibid.* **13**, 727-742 (1985).

36. M. Aerts, M. Jacobs, J.-P. Hernalsteens, M. van Montagu and J. Schell, *Plant Sci. Lett.* **17**, 43-50 (1979).

37. D. Botstein, R.L. White, M. Skolnick and R.W. Davis, *Am. J. Hum. Genet.* **32**, 314-331 (1980).

38. The sources of the genome measurements are: *Arabidopsis*, (24); Mung Bean, M.G. Murray, J.D. Palmer and W.F. Thompson, *Biochemistry* **18**, 5259-5266 (1979); Cotton, V. Walbot and L.S. Dure III, *J. Mol. Biol.* **101**, 503-536 (1976); Tobacco, J.L. Zimmerman and R.B. Goldberg, *Chromosoma* (Berl.) **59**, 227-252 (1977); Soybean, R.B. Goldberg, *Biochem. Genet.* **16**, 45-68 (1978); Pea, M.G. Murray, R.E. Cuellar and W.F. Thompson, *Biochemistry* **17**, 5781-5790 (1978); Wheat, D.B. Smith and R.B. Flavell, *Chromosoma* (Berl.) **50**, 223-242 (1975) and R.B. Flavell and D.B. Smith, *Heredity* **37**, 231-252 (1976).

39. We would like to thank the other members of the Meyerowitz lab for comments on the manuscript. Our *Arabidopsis* work is supported by National Science Foundation grant PCM-8408504 to E.M.M.

Chapter 2

Characterization of the Genome of *Arabidopsis thaliana*

(in press in the Journal of Molecular Biology)

Summary

The small crucifer *Arabidopsis thaliana* has many useful features as an experimental organism for the study of plant molecular biology. It has a four-week life cycle, only five chromosomes and a genome size less than half that of *Drosophila*. To characterize the DNA sequence organization of this plant, we have randomly selected 50 recombinant lambda clones containing inserts with an average length of 12.8 kb and analyzed their content of repetitive and unique DNA by various genome blot, restriction digestion and RNA blot procedures. The conclusions that can be drawn include:

1) The DNA represented in this random sample is composed predominantly of single-copy sequences. This presumably reflects the organization of the *Arabidopsis* genome as a whole and supports prior conclusions reached on the basis of kinetics of DNA reassociation.

2) The DNA which encodes the ribosomal RNAs constitutes the only major class of cloned nuclear repetitive DNA. It consists of approximately 570 tandem copies of a heterogeneous 9.9 kb repeat unit.

3) There are an average of approximately 660 copies of the chloroplast genome per cell. Therefore, the chloroplast genome constitutes the major component of the repetitive sequences found in *Arabidopsis thaliana* DNA made from whole plants.

4) The inner cytosine in the sequence CCGG is methylated more often than the outer cytosine in the tandem ribosomal DNA units, whereas very few differences in the methylation state of these two cytosines are detected in unique sequences.

1. Introduction

The genomes of higher plants exhibit great variation both in sequence content and sequence organization. The nuclear DNA content of the angiosperms varies over nearly three orders of magnitude (Bennett & Smith, 1976) and the percentage of the nuclear genome present in repeated sequences varies from the small quantity found in *Arabidopsis* (Leutwiler *et al.*, 1984) to 70-75% found in wheat, rye and pea (Flavell & Smith, 1976; Smith & Flavell, 1977; Murray *et al.*, 1978). The sequence interspersion patterns also vary greatly, even among closely related plants. This is demonstrated by the genome of the pea (*Pisum sativum L.*), which apparently contains no single-copy sequences longer than 1000 nucleotides (Murray *et al.*, 1978), and the genome of the mung bean (*Vigna radiata*; a member of the same subfamily), in which almost 50% of the single-copy sequences are present in blocks greater than 6700 nucleotides in length (Murray *et al.*, 1979). Genome size, repetitive sequence content and sequence interspersion are interrelated: larger genomes tend to contain a greater percentage of repetitive sequences and have more of their single-copy sequences contained in regions of short period interspersion (Flavell, 1980).

Leutwiler *et al.* (1984) have performed a kinetic analysis on total *Arabidopsis* DNA and found it to be composed of 10-14% rapidly reannealing sequences, 23-27% middle repetitive sequences and 50-55% single-copy sequences. In addition, they have determined that the majority of the middle repetitive DNA is derived from the chloroplast genome and have calculated that the haploid *Arabidopsis* nuclear genome contains only 7×10^7 base pairs, making it the smallest higher plant genome characterized to date.

Two different methods have been used to examine sequence interspersion patterns in different organisms: kinetic analysis of DNA reassociation (Davidson *et al.*, 1973; Graham *et al.*, 1974) and examination of reannealed repetitive DNA

in the electron microscope (Manning *et al.*, 1975; Walbot & Dure, 1976). In the first method the kinetics of reassociation of DNA fragments sheared to various lengths is determined. By comparing the quantity of rapidly renaturing DNA in fragments of different sizes, information about the sequence interspersion pattern is obtained. In the second method, DNA is denatured, allowed to reassociate and the duplex molecules examined in the electron microscope. This type of experiment allows direct determination of the distribution of lengths of repetitive sequences and an estimate of the minimum length between repeated sequences. Both of these methods characterize the genome in a relatively unbiased way.

We have used a third method to analyze sequence interspersion in the *Arabidopsis* genome. We have characterized the unique and repetitive sequences found in randomly chosen recombinant lambda clones. While this method introduces some bias into the analysis because of the possibility that the clones fail to represent some regions of the genome, it offers the advantage of allowing much more detailed characterization of repetitive sequences than the other methods. Our analysis of 0.8% of the *Arabidopsis* genome shows that there is very little repetitive DNA in the *Arabidopsis* nucleus and that the single-copy DNA is present in extremely long contiguous blocks. In addition, we show that most of the repetitive DNA which we have cloned from the *Arabidopsis* nucleus is ribosomal DNA and we present a description of that DNA.

2. Materials and Methods

(a) Plant culture

The *Arabidopsis* strains used in this work were: Columbia, obtained from A. Kleinhofs, Program in Genetics, Washington State University, Pullman, WA 99164; Landsberg *erecta*, obtained from F. J. Braaksma, Department of Genetics, Biology Centre, Haren, The Netherlands; and Niederzenz, obtained from

A. R. Kranz, Botanisches Institut, J. W. Goethe- Universitat, Frankfurt am Main, Federal Republic of Germany. Landsberg and Niederzenz are two different wild type strains of *Arabidopsis thaliana* which were isolated in Landsberg, Germany and Niederzenz, Germany respectively (Röbbelen, 1965; Kranz, 1978). The Landsberg strain used in this work bears a homozygous recessive mutation, *erecta*, which makes the plants more compact and easier to culture in large numbers. The Columbia strain was derived from the Landsberg strain by Redei (1970). Plants were grown on a mixture of sterile soil, peat moss and sand (3:3:1, v/v) under constant illumination (7000 lux) at 25°C and 70% relative humidity.

Axenic cultures of *Arabidopsis* were prepared by surface sterilizing ~50 seeds in 5% sodium hypochlorite for 8 min, followed by three rinses in an excess of sterile distilled water. The sterilized seeds were then introduced into 100 ml liquid cultures of the medium of Redei (1965). The liquid cultures were grown under the same conditions of light and temperature as the soil cultures.

(b) General DNA and recombinant DNA techniques

Restriction digestions were performed as described by Davis *et al.* (1980). ^{32}P-labeled DNA was prepared by nick translation following the method of Rigby *et al.* (1977). Genomic libraries were constructed using DNA from the Columbia strain as described by Meyerowitz and Martin (1984). One of the two libraries was amplified before use according to the method described by Maniatis *et al.* (1982). Lambda clones with numbers in the range of 000-099 came from the unamplified library, while those with numbers in the range 100-199 came from the amplified library. The clone nomenclature system has been described (Meyerowitz & Martin, 1984).

National Institutes of Health guidelines were followed for the P1-EK1 level containment of recombinant DNA-bearing organisms.

(c) Nucleic acid preparations

Plasmid and bacteriophage DNA preparations were performed as described by Davis *et al.* (1980).

DNA was extracted from 1.0-1.5 g of whole plants by grinding with 0.5 g of glass beads (75-150 µm, Sigma) in a mortar containing 3 ml of 0.2 M Tris-HCl pH 8.0, 0.1 M EDTA, 1% sodium N-lauroyl sarcosine and 100 µg/ml proteinase K (Merck). This was incubated at 47°C for 1 h followed by centrifugation in a table-top centrifuge. The supernatant was precipitated with 6 ml ethanol and centrifuged at 10,000 rpm for 15 min in an SS-34 rotor. The pellet was resuspended in 3 ml of TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA) and ethanol-precipitated again. The second pellet was resuspended in 4 ml of TE, and 4.5 g of CsCl and 400 µl of 10 mg/ml ethidium bromide were added. This solution was centrifuged at 53,000 rpm for 18-20 h in a VTi65 rotor. Following centrifugation the UV fluorescent band was removed with a syringe, diluted to 2 ml with distilled water and extracted twice with 4 ml of n-butanol to remove the ethidium bromide. The DNA was then precipitated by addition of two volumes of ethanol followed by centrifugation at 10,000 rpm for 15 min in an SS-34 rotor. The pellet was washed with 70% ethanol, air-dried and resuspended in 100 µl of TE overnight at 4°C.

RNA was extracted from 0.25-0.5 g of whole plants by grinding with 0.25 g of glass beads in a mortar containing 1 ml of 100 mM Tris-HCl pH 8.5, 100 mM NaCl, 20 mM EDTA and 1% (w/v) sodium N-lauroyl sarcosine. The homogenate was extracted twice with phenol/chloroform (1:1, v/v) and once with chloroform followed by ethanol precipitation and centrifugation at 10,000 rpm for 15 min in an SS-34 rotor. The pellet was washed in 70% ethanol, air-dried and resuspended in 50 µl of TE.

Chloroplast DNA was prepared as described by Leutwiler *et al.* (1984)

except that the DNA was not purified by CsCl-ethidium bromide centrifugation.

(d) Gel electrophoresis

Agarose gel electrophoresis of DNA and RNA was performed as described by Meyerowitz and Martin (1984). Renaturation of RNA in formaldehyde-agarose gels was performed by extensive washing of the gel in distilled water, then in 0.1 M ammonium acetate and finally staining in 0.5 μg/ml ethidium bromide. Restriction fragment sizes of λcI857 S7 are from the sequence of Sanger *et al.* (1982).

(e) Filter binding and hybridization of nucleic acids

DNA and RNA were transferred from agarose gels as described by Meyerowitz and Martin (1984).

Dot blot filters were prepared using a 96-well manifold (Bio-Rad). A series of 12 twofold serial dilutions of *Arabidopsis* genomic DNA was prepared with mouse DNA of concentration equal to the original *Arabidopsis* DNA concentration, thus maintaining the overall DNA concentration as the *Arabidopsis* DNA concentration decreased from 160 μg/ml to 78 ng/ml. Sixty μl of each solution was mixed with 180 μl of 0.4 M NaOH, incubated at room temperature for 10 min and then chilled on ice. Two hundred forty μl of 2 M ammonium acetate was added and 50 μl of the resulting solution was filtered through a nitrocellulose filter equilibrated in 1 M ammonium acetate in each of eight wells in the manifold. All of the wells were rinsed with 200 μl of 1 M ammonium acetate; the filter was cut into eight identical strips containing 12 dots each and baked for 2 h at 80°C in a vacuum oven.

All hybridizations were in 50% formamide, 5X SSPE (1X=180 mM NaCl, 10 mM NaH_2PO_4, 8 mM NaOH, 1 mM Na_2EDTA pH 7.0 Davis *et al.*, 1980), 100 μg/ml sonicated and denatured salmon testis DNA, 1X Denhardt's solution

(0.02% *Ficoll*, 0.02% polyvinylpyrrolidone, 0.02% bovine serum albumin, Denhardt, 1966) and 0.1% sodium dodecyl sulfate (SDS) at 43°C. After hybridization, filters were washed in 1X SSPE, 0.1% SDS at room temperature except for the dot blot filters which were washed at 43°C. Hybridized DNA was removed by washing filters in 0.01X SSPE, 0.1% SDS at 100°C.

(f) Quantitation of dot blot data

After washing, the individual dots were separated into 1 cm squares of nitrocellulose, dissolved in liquifluor/toluene scintillation cocktail and counted for 50 min (or to 1% accuracy) in a Beckman LS-250 liquid scintillation counter. The data from the duplicate filters were averaged and the cpm plotted versus the quantity of DNA on the filter. Lines were fitted by linear regression. The slopes of the lines were then converted to μmoles of clone hybridized/μg genomic DNA bound to the filter by dividing by the specific activity of the clone (cpm/μg) and then dividing by the molecular weight of the insert portion of each lambda clone used as a probe. The resulting values can be compared directly to give relative copy numbers in the genomic DNA.

3. Results

(a) Analysis of random cloned fragments by reverse genome blotting

In order to characterize the sequence interspersion patterns of the *Arabidopsis* genome, we examined 50 randomly selected lambda recombinant clones. Clones were selected from two different *Eco*RI partial digest libraries which were made in the lambda vector Sep6 (Meyerowitz & Hogness, 1982), and contained DNA from whole plants of the Columbia strain of *Arabidopsis*. Purified cloned DNA was digested with *Eco*RI, fractionated by agarose gel electrophoresis, and blotted to nitrocellulose filters. These filters were then hybridized to ^{32}P-

labeled *Arabidopsis* genomic DNA derived from the Landsberg strain, and labeled by nick translation. The degree of hybridization, as revealed by autoradiograms of these filters, allowed the assignment of the clones to two classes (high and low) reflecting the repetition frequency of their *Arabidopsis* DNA inserts in the genomic DNA. A typical blot is shown in Figure 1, with λbAt113 falling in the highly repeated class, and the others in the low repetition group. As can be seen in Table 1, the majority of the clones fall into the low repetition class.

The high repetition frequency class could be subdivided into three groups of clones based on the *Eco*RI restriction fragments contained in the *Arabidopsis* DNA insert and the pattern of hybridization to those fragments. Eight of the 15 repetitive clones (λbAt002, λbAt025, λbAt106, λbAt107, λbAt113, λbAt122, λbAt123 and λbAt125) contain a set of four identical *Eco*RI restriction fragments and a fifth fragment which displays some variation in length. All of the restriction fragments in this group of clones show intense hybridization to the genomic DNA probe. This set of clones will be referred to below as subdivision I of the high repetition frequency class.

For four of the remaining clones (λbAt003, λbAt102, λbAt112 and λbAt124; subdivision II) the genomic DNA probe showed intense hybridization to all of the restriction fragments present in the *Eco*RI-digested clone DNA, while the final three (λbAt006, λbAt010 and λbAt026 subdivision III) showed strong hybridization to only a subset of their fragments. λbAt006 contains two *Eco*RI fragments of 6.4 kilobase-pairs (kb) as well as one of 1.0 kb. When *Eco*RI-digested DNA from the clone was hybridized with labeled genomic DNA, strong hybridization was observed to one of the 6.4 kb bands while the other two fragments appeared to be low copy number. λbAt010 contains five *Eco*RI restriction fragments of 8.9, 1.3, 1.25, 0.8 and 0.65 kb. Of these the 0.8 kb and the 0.65 kb fragments appear to contain no repetitive sequences, while the remaining fragments all appear to contain some repetitive DNA. λbAt026 contains *Eco*RI fragments of 6.1, 5.3 and

2.9 kb. When labeled genomic DNA was hybridized to this clone only the 6.1 kb *Eco*RI fragment appeared to be repetitive.

In the following, the classes of cloned sequences will be treated separately, the low repetition class first followed by each subdivision of the repetitive clones.

(b) Further analysis of the low repetition frequency class

The next experiment was designed to determine if the sequences represented in the recombinant clones were present exclusively on the same size restriction fragments in the genome of *Arabidopsis.* Genomic DNA from the Landsberg *erecta* strain was digested with *Eco*RI, fractionated by agarose gel electrophoresis and blotted to nitrocellulose filters. ^{32}P-labeled DNA from each clone in the low copy number class was hybridized individually to a filter. All but five of these clones hybridized only to restriction fragments which were the same size as the restriction fragments contained in the clone. In addition, with the exception of λbAt119 which is discussed below, the autoradiographic signals on these filters were of approximately the same intensity.

To determine the approximate copy number of these clones in the *Arabidopsis* genome, a reconstruction experiment was performed using two representative clones: λbAt114 and λbAt115. *Arabidopsis* genomic DNA was digested in parallel with digests containing the equivalent of one, three or ten copies per haploid genome of clone DNA assuming a total cellular genome size of 1×10^8 base pairs (Leutwiler *et al.*, 1984). The digestion products were fractionated by agarose gel electrophoresis, denatured and transferred to a nitrocellulose filter. Each filter was allowed to hybridize with ^{32}P-labeled DNA from the same clone used to make the filter. Autoradiograms of these filters demonstrate that each clone hybridizes with approximately equal intensity to the genomic DNA and the DNA representing the single-copy reconstruction (Fig. 2).

Of the five clones which hybridized to genomic restriction fragments which

were of *different* sizes than those contained in the cloned DNA, four clones hybridized to patterns of restriction fragments which can be accounted for by restriction site polymorphism. λbAt047 contains *Eco*RI fragments of 6.7, 3.5 and 3.4 kb in length. When used as a hybridization probe of *Eco*RI-digested genomic DNA this clone hybridized to fragments of 6.7, 3.5, 1.35, 1.3 and 0.7 kb. λbAt132 contains three *Eco*RI fragments of *Arabidopsis* DNA which are 8.1, 4.7 and 1.2 kb in length. Hybridization to *Eco*RI-digested genomic DNA revealed fragments of 6.2, 4.7, 1.9 and 1.2 kb. λbAt133 contains four *Eco*RI fragments in its insert which are 7.9, 7.1, 3.7, and 1.4 kb in length. The hybridization pattern of this clone to *Eco*RI-digested genomic DNA revealed hybridization to 11.6, 7.1 and 1.4 kb fragments. For each of these three clones most of the fragments are conserved between the clone and the genomic DNA and the sum of the lengths of the fragments seen in the blot pattern is the same as the overall length of the clone insert.

λbAt108 contains three *Eco*RI restriction fragments which are 4.9, 4.4 and 1.0 kb in length. When used as a hybridization probe against *Eco*RI-digested genomic DNA this clone hybridized to fragments of 4.4 and 1.0 kb only. When hybridized to *Eco*RI-digested genomic DNA of a different ecotype (Niederzenz) hybridization to fragments of 4.9, 4.4 and 1.0 kb was detected. If the restriction fragment found in the Landsberg strain homologous to the cloned 4.9 kb fragment were large enough to approach the mean size of the DNA isolated it would no longer run as a discrete band on the agarose gel.

The polymorphisms in these four clones are not unexpected, since the recombinant libraries were constructed with DNA from the Columbia strain while the blots were performed with DNA from the closely related, but not identical, Landsberg *erecta* strain. The existence of polymorphisms of the type described confirms that the polymorphic clones, and those hybridizing to genomic DNA with identical intensity, do indeed represent DNA sequences present only once per

haploid genome in the plant.

Two clones hybridized to other restriction fragments in addition to those contained in the clone. In each case the intensity of the hybridization signal of the new restriction fragments was approximately equal to the hybridization signal of the fragments the same size as those contained in the clone. λbAt121 contains *Eco*RI fragments of 7.1 and 2.2 kb. When DNA from this clone was ^{32}P-labeled and hybridized to *Eco*RI-digested genomic DNA, hybridization to fragments of 7.1, 6.1, 5.1, 3.6 and 2.2 kb was observed. The insert of λbAt131 consists of a single *Eco*RI fragment of 11.0 kb. Hybridization to *Eco*RI-digested genomic DNA revealed hybridization to an 11.0 kb fragment and a 3.0 kb fragment. This hybridization to extra fragments may represent duplicated genomic copies of the sequences contained in the clones, or heterozygosity for restriction fragment length polymorphisms in the plants from which the DNA used in the genomic blots was extracted.

One clone (λbAt119) hybridized only to restriction fragments which were the same size as those contained in the clone but the signal intensity was severalfold greater than with the other clones in the low copy number class. In order to determine the approximate number of copies of this sequence per haploid *Arabidopsis* genome, a reconstruction experiment similar to that described above was performed using λbAt119. An autoradiogram of the filter resulting from this experiment revealed that λbAt119 hybridizes more intensely to the genomic DNA than to the DNA sample containing the equivalent of ten copies per haploid genome.

Because of the greater degree of hybridization as compared to the other low copy number clones, it was possible that this clone represented a bacterial contaminant present in the plant DNA. To test this hypothesis DNA was prepared from *Arabidopsis* plants that were grown under axenic conditions. This DNA and

Arabidopsis DNA made from plants grown on soil were digested with *Eco*RI, fractionated by agarose gel electrophoresis and transferred to a nitrocellulose filter. When this filter was hybridized with ^{32}P-labeled DNA from λbAt119 and autoradiographed it revealed an approximately equal degree of hybridization to each DNA sample. Therefore, this clone does represent sequences present in the plant genome, but present in several indistinguishable copies.

(c) The clones of repetitive subdivision I contain Arabidopsis rDNA sequences

Since all of the clones of subdivision I appeared to be similar, they clearly constituted some major family of repetitive DNA. Five of the eight clones (λbAt002, λbAt025, λbAt106, λbAt122 and λbAt125) contained the same five restriction fragments which had lengths of 3.75, 2.35, 1.65, 1.5 and 0.65 kb. The remaining clones lacked the 1.65 kb *Eco*RI fragment and contained one additional *Eco*RI fragment: λbAt107 contained a fragment of 1.9 kb, λbAt113 a fragment of 1.6 kb and λbAt123 a fragment of 2.0 kb. To determine if all of these clones contained homologous sequences, approximately equal quantities of the cloned DNAs were digested with *Eco*RI, fractionated by agarose gel electrophoresis and transferred to a nitrocellulose filter. ^{32}P-labeled λbAt002 DNA was then allowed to hybridize to this filter. Autoradiography of the filter revealed that all of the fragments in all of the subdivision I repetitive clones were hybridized by this probe. When the vector λSep6 was used as the probe, none of the insert fragments in any of the clones hybridized to the labeled DNA.

To determine if the heterogeneity observed in the length of the 1.65 kb *Eco*RI fragment was due to the presence of fragments of various lengths in the genome, the following experiment was performed. Genomic DNA was digested with *Eco*RI, fractionated on an agarose gel and transferred to nitrocellulose. A plasmid subclone containing the 1.65 kb *Eco*RI fragment from λbAt002 was ^{32}P-

labeled by nick translation and used as a hybridization probe on this filter. Autoradiography of the filter showed that this fragment hybridizes strongly to genomic *Eco*RI fragments of 2.35, 1.65 and 1.6 kb and also hybridizes weakly to fragments of many other sizes.

In all of the clones some of the fragments were present in greater molar quantities than others. This suggested a tandem repeat type organization and the 9.9 kb length of the repeat unit suggested that these units might be rRNA coding regions. To test this hypothesis total *Arabidopsis* RNA was prepared, denatured and fractionated on a formaldehyde-agarose gel. One lane of this gel was blotted to a nitrocellulose filter, while another was extensively washed, the RNA allowed to renature and stained with ethidium bromide. The filter was allowed to hybridize with ^{32}P-labeled λbAt002 DNA and the location of hybridization determined by autoradiography. Hybridization was observed at two locations which corresponded in migration with the two major bands observed in the ethidium-stained portion of the gel. Thus, in all probability this clone contains sequences which code for the ribosomal RNAs of *Arabidopsis*.

(d) The organization of the rDNA repeat unit is typical of eukaryotic organisms

To determine if all of the cloned 9.9 kb rDNA repeat units are the same, the restriction endonuclease maps for several enzymes were determined for two independent clones (λbAt002 and λbAt025). The repeat units contained in these clones have identical maps for *Eco*RI, *Bgl*II, *Xho*I and *Xba*I. The restriction mapping also confirmed that the units are organized as tandem repeats with each of these clones containing 1.5 repeats. The restriction map of the repeat unit is presented in Figure 3. One 700 bp region of the repeat unit is cleaved into fragments smaller than 100 bp by *Sal*I (Fig. 3). This region is partly contained in the *Eco*RI restriction fragment (the right-most in the figure) which displays

heterogeneity in length in the different clones.

The location of the coding regions within the repeat unit was determined by further RNA blot analysis. Subclones containing the five *Eco*RI fragments of the repeat unit were ^{32}P-labeled by nick translation and then used as hybridization probes against total RNA filters prepared as described above. The hybridization patterns of each of the subclones is presented in Figure 4. As can be seen in the figure the hybridization data can be assembled into a consistent map of the regions of the clone coding for each of the large ribosomal RNAs.

(e) There are approximately 570 copies of the rDNA repeat unit in Arabidopsis

To determine the approximate copy number of the rDNA repeat unit in the haploid genome, a quantitative dot blot procedure was used. Three sets of two filters, each of which consisted of a series of dots of serial dilutions of *Arabidopsis* genomic DNA, were hybridized with ^{32}P-labeled λbAt002, λbAt114 and λbAt115. After hybridization the dots were cut apart and the extent of hybridization to each dot was determined by liquid scintillation spectrometry. The two values for each DNA concentration were averaged and lines fitted to the data by linear regression (Fig. 5). All points were used for the two unique clones but the three highest DNA concentrations were omitted for the rDNA clone where the curve was obviously non-linear. The slopes of these lines represent cpm bound/μg of DNA. These can be compared directly after correcting for the specific activity of the different probes and the length of the segment being hybridized in the genomic DNA. After these corrections are made, the single-copy clones λbAt114 and λbAt115 have 2.2×10^{-13} and 2.4×10^{-13} μmoles probe/μg genomic DNA bound to the filters. By comparison λbAt002 has 1.3×10^{-10} μmoles probe/μg genomic DNA bound to the filter. By division of these two values it can be calculated that there are approximately 570 copies of the rDNA repeat per

haploid genome in *Arabidopsis*.

(f) Arabidopsis rDNA and unique sequences both contain 5-methyl cytosine

To examine the relative levels of methylation of the *Arabidopsis* rDNA and unique sequences, DNA blots with methylation-sensitive restriction endonucleases were performed. Genomic DNA was digested in parallel with *Hpa*II and *Msp*I, which cut, respectively, at CCGG/^{me}CCGG and CCGG/C^{me}CGG sequences, fractionated by agarose gel electrophoresis and transferred to nitrocellulose. Eight filters were prepared and each was hybridized with ^{32}P-labeled DNA from a different lambda clone containing unique sequences. The clones used were λbAt103, λbAt104, λbAt105, λbAt114, λbAt115, λbAt116, λbAt117 and λbAt118. Autoradiograms of these filters revealed that the hybridization patterns of these clones to genomic DNA digested with *Hpa*II or *Msp*I were very similar (Fig. 6). Pooling the data from all eight clones, there were 38 bands present in both digests, five which were present only in the *Msp*I digests and four which were present in only the *Hpa*II digest. Following this experiment, the hybridized DNA was removed from one of these filters and the filter allowed to hybridize with ^{32}P-labeled DNA from one of the rDNA clones (λbAt025). The hybridization pattern of this clone demonstrates that the *Arabidopsis* rDNA is cleaved by both *Hpa*II and *Msp*I, but that the DNA is reduced to smaller fragments by *Msp*I (Fig. 6). Therefore, the inner cytosine in the sequence CCGG is more frequently methylated than the outer cytosine in *Arabidopsis* rDNA sequences.

(g) Repetitive subdivision II is composed of clones containing chloroplast DNA

The clones of subdivision II, which show equal hybridization of genomic DNA to all fragments, all show approximately the same intensity of hybridization of the ^{32}P-labeled genomic DNA. One of these clones (λbAt003) has been shown

to hybridize to partially purified chloroplast DNA and presumably contains chloroplast sequences (Leutwiler *et al.*, 1984). Because of the similar apparent copy number in the genomic DNA it was considered likely that all of these represent chloroplast clones. To determine if this is true, DNA from each of these clones and DNA from λbAt025, a ribosomal DNA clone, was digested with *Eco*RI, fractionated by agarose gel electrophoresis and transferred to nitrocellulose. Two filters were made in parallel: one was hybridized with ^{32}P-labeled genomic DNA while the other was allowed to hybridize with ^{32}P-labeled purified chloroplast DNA. As can be seen for three of the clones in Figure 7, the genomic DNA hybridizes with approximately equal intensity to the rDNA clone and to each of the other clones, whereas the purified chloroplast DNA hybridizes only slightly to the rDNA clone and strongly to each of the other clones. This evidence demonstrates that each of the clones in this class contains DNA which is a part of the chloroplast genome.

To determine the approximate number of chloroplast genomes per haploid genome in the nucleus, a quantitative dot blot experiment similar to that described earlier for the rDNA, but using the chloroplast clone λbAt003, was performed. As was the case for the rDNA the points representing the highest concentrations of genomic DNA were not included in the linear regression calculations due to their obvious non-linearity (Fig. 5). After correction for specific activity and length of the hybridizing fragment the chloroplast clone was found to have bound to the filter at a level of 7.7×10^{-11} μmoles probe/μg genomic DNA. Compared to the average of the unique clones (2.3×10^{-13} μmoles probe/μg genomic DNA) this gives an approximate copy number of 330 copies per haploid genome or an average of 660 copies per cell in whole plants. Because the chloroplast genome may vary in copy number with respect to the nuclear genome, and because the DNA used in this experiment came from various cell types, this

number represents only an estimate of the chloroplast genome copy number.

(h) The repetitive subdivision III clones can be subdivided into two groups on the basis of hybridization to genomic DNA

The three remaining middle repetitive clones (λbAt006, λbAt010 and λbAt026) were distinctive in having genomic DNA hybridize to a greater degree to some of their restriction fragments than to others. As a means of further characterization these clones were labeled with ^{32}P and used as probes of *Eco*RI-digested *Arabidopsis* genomic DNA. When λbAt006, which contains two 6.4 kb and one 1.0 kb *Eco*RI fragments, was used as a hybridization probe, strong hybridization was observed to fragments of 6.4 and 2.7 kb while weaker hybridization was observed to a 1.0 kb fragment. Thus, it appears that the repetitive 6.4 kb *Eco*RI fragment detected by the reverse genome blot hybridizes to genomic DNA fragments of 6.4 kb and 2.7 kb.

The hybridization result suggested that λbAt006 might in fact be the result of the ligation of a highly repetitive 6.4 kb fragment with two unique fragments during the construction of the clone library. To test this possibility genomic DNA and DNA from the clone were digested with *Bam*HI and *Xho*I. This DNA was fractionated on an agarose gel and transferred to nitrocellulose. This filter was allowed to hybridize with ^{32}P-labeled DNA of the 1.0 kb *Eco*RI restriction fragment which had been purified on an agarose gel. Autoradiography of this hybridization revealed that the probe had hybridized to a 2.2 kb *Bam*HI fragment and a 6.7 kb *Xho*I fragment in both the cloned and genomic DNAs. Since the 1.0 kb *Eco*RI fragment is located between the two 6.4 kb fragments and does not contain sites for either *Bam*HI or *Xho*I, the fragments to which this probe hybridizes must span both internal *Eco*RI sites. Thus, this clone is unlikely to be the result of a cloning artifact and represents a junction between repetitive and

unique sequences in the *Arabidopsis* genome.

When either λbAt010 or λbAt026 was hybridized to genomic DNA a similar although not identical pattern was observed (Fig. 8). Both probes hybridized to many fragments and in particular showed strong hybridization to DNA of sizes 6.1 and 4.5 kb. Therefore, both of these clones contain repetitive sequences which are present many times in the genome and are contained on *Eco*RI fragments of many different sizes.

(i) Localization of the repetitive sequences in the clones λbAt010 and λbAt026

In order to determine the location of repetitive sequences in these two lambda clones it was first necessary to determine their restriction maps. Sites of cleavage for the restriction enzymes *Bam*HI, *Eco*RI, *Sal*I and *Xba*I were determined for λbAt010 while cleavage sites for *Bgl*II, *Eco*RI, *Hin*dIII and *Sal*I were determined for λbAt026. The resulting restriction maps are presented in Figure 9.

DNA from the clone λbAt010 was digested with *Bam*HI, *Xba*I and *Eco*RI + *Sal*I. This DNA was fractionated by agarose gel electrophoresis and transferred to a nitrocellulose filter. The gel used did not allow visualization of fragments smaller than approximately 0.7 kb. This filter was then hybridized with ^{32}P-labeled *Arabidopsis* genomic DNA. Autoradiography of this filter revealed intense hybridization to all fragments of DNA derived from the insert of the clone and large enough to be detected with the exception of the 1.3 and 1.25 kb *Xba*I fragments, the 0.8 kb *Eco*RI fragment and the 1.6 kb *Eco*RI-*Sal*I fragment. Previous blotting experiments had shown that intense hybridization was not observed at the position of the 0.65 kb *Eco*RI fragment. These data indicate that there are at least four regions of repetitive sequences in λbAt010 which are separated by sequences present in fewer copies in the *Arabidopsis* genome (Fig. 9).

A similar gel blotting experiment was performed to determine what region of the clone λbAt026 contained repetitive sequences. A gel blot filter was prepared with lanes containing λbAt026 digested with *Bgl*II + *Sal*I and *Hin*dIII. This filter was hybridized with ^{32}P-labeled *Arabidopsis* genomic DNA. Autoradiography revealed strong hybridization which was limited to the 1.7 kb *Bgl*II-*Sal*I fragment and to the 6.0 kb *Hin*dIII fragment. These data indicate that the highly repeated sequences in λbAt026 are confined to the 1.4 kb *Hin*dIII-*Bgl*II fragment (Fig. 9).

(j) The repetitive elements in λbAt010 and λbAt026 do not cross-hybridize

To determine if the elements contained in λbAt010 and λbAt026 are related, an experiment was performed to detect homologous sequences present in the repetitive regions of these two clones. λbAt010 DNA was digested with *Eco*RI and *Bam*HI, fractionated on an agarose gel and transferred to a nitrocellulose filter. A plasmid subclone containing the 6.1 kb *Eco*RI fragment from λbAt026 was ^{32}P-labeled and used as a hybridization probe. Autoradiography revealed that this probe, which contains the repetitive sequences found in λbAt026, does not hybridize to any of the insert fragments contained in λbAt010.

4. Discussion

The work described was undertaken in order to obtain a clear picture of the organization of the genome of *Arabidopsis thaliana*. It was known from the work of Leutwiler *et al.* (1984) that the *Arabidopsis* genome is extremely small and that it contains a high proportion of presumed single-copy sequences. In order to determine the nature of the repetitive sequences and the pattern of their interspersion among the unique sequences we have analyzed 50 randomly chosen lambda clones containing a total of greater than 600 kb of DNA. The validity of

this analysis is dependent on three conditions. First, the recombinant libraries used must be representative of the genome. Because our libraries were constructed by *Eco*RI partial digestion it is clear that we have excluded any *Eco*RI fragments which exceed the capacity of the λ-vector. We may also be selecting against regions of the genome containing many closely spaced *Eco*RI sites if we have overdigested the DNA used to make the libraries (Seed *et al.*, 1982); however, in the construction of the library precautions were taken to avoid this possibility. In addition, our libraries may exclude some regions of the genome containing inverted repeats which render recombinant λ-phage containing them inviable on rec^+ hosts (Leach and Stahl, 1983; Wyman *et al.*, 1985). Not all inverted repeats behave in this manner since we have previously cloned such a repeat using this vector and host (Meyerowitz & Hogness, 1982). One test indicates that at least the amplified library is indeed representative. We have screened this library for four different genes using heterologous probes and have always obtained phage containing the sequences of interest at approximately the expected frequency (L. S. Leutwiler, R. E. Pruitt and C. Chang, unpublished). The second condition which must be met is that hybridization should be performed under conditions allowing the recognition of small or partially conserved repetitive sequences. As described in Materials and Methods, all hybridizations and filter washes were carried out under non-stringent conditions (Tm-25°C). Finally, the sample of clones examined must be large enough to be representative of the entire genome. The small genome of *Arabidopsis* makes it relatively easy to look at a reasonable proportion of the nuclear sequences. We have examined 578 kb of nuclear DNA which, using the haploid nuclear genome value of 7×10^4 kb given by Leutwiler *et al.* (1984), constitutes approximately 0.8% of the nuclear genome.

Of the 50 clones examined, four contain chloroplast sequences. The 46

non-chloroplast clones can be divided into five groups based on their DNA sequence content. The largest group is made up of 32 of the 50 clones, which contain only sequences which appear to be present once in the haploid *Arabidopsis* genome. Four of these clones contain *Eco*RI sites which are polymorphic between the two closely related ecotypes Columbia and Landsberg. Restriction site polymorphisms between Landsberg and the more distantly related strain Niederzenz can be found in many of these clones and will form the basis for a restriction fragment length polymorphism (RFLP) genetic map now being constructed (R. E. Pruitt, work in progress). The ease with which RFLPs can be detected and the nature of the polymorphisms provide additional evidence that these clones do represent sequences which are unique in the genome.

The second group consists of two clones (λbAt121 and λbAt131) which appear to contain sequences which occur in a small number of discrete locations in the genome. A specific example of such a sequence is known in *Arabidopsis*; there are three distinct genes encoding the chlorophyll a/b binding protein, all located in a small gene cluster (L. S. Leutwiler and E. M. Meyerowitz, unpublished). Another possibility is that the extra bands seen in the genome blots with these two clones represent RFLPs which are segregating in the Landsberg population. Because *Arabidopsis* normally is self-pollinating, all sequences in an individual plant will tend to become homozygous. Because our Landsberg population was started from a single plant only three generations ago, it is possible that polymorphisms are present in our population if the original plant was heterozygous for such polymorphisms.

The third group consists of only one clone, λbAt119. This clone contains a DNA segment which is present at a level exceeding ten copies per haploid genome and is highly conserved over its entire length. Sequences homologous with those in the clone were found to be present in the DNA of plants grown from sterilized

seeds under axenic conditions, ruling out the possibility that the cloned DNA was that of a contaminating soil bacterium or fungus. It is possible that this clone contains mitochondrial DNA sequences; the copy number estimated from a copy number reconstruction experiment is consistent with the quantity of mitochondrial DNA expected in *Arabidopsis* (Leutwiler *et al.*, 1984).

The fourth group of clones (λbAt006, λbAt010 and λbAt026) contains both unique and repetitive sequences. Two of the clones contain repetitive sequences which do not contain a well-conserved *Eco*RI fragment while the repeat unit found in the other clone does contain a well-conserved *Eco*RI fragment. λbAt006 and λbAt026 each contain long contiguous blocks of unique sequences, indicating that these repeat sequences are interspersed among the predominant unique sequences. The exact nature of these repetitive families has not been determined.

The last group of clones are those containing rDNA sequences. Analysis of these clones indicates that the rDNA of *Arabidopsis* is typical of the rDNA of plants. In higher plants the number of rDNA repeats per cell varies from 1,000 to greater than 30,000 (Ingle *et al.*, 1975) and the length of the repeats varies from 8 to 12 kb (Leweke & Hemleben, 1982). *Arabidopsis* falls within both of these ranges having approximately 1150 copies per cell of a 9.9 kb repeat unit. It is not known if these repeats are present in one or several clusters within the genome. Cytologically there are two nucleolus organizers, a large one on chromosome 4 and a smaller one on chromosome 2 (Sears & Lee-Chen, 1970). Therefore, it seems likely that the rDNA repeats are located in two clusters on these chromosomes.

The *Arabidopsis* rDNA repeat units are heterogeneous in length. In every variant repeat unit cloned, the length difference resides in the *Eco*RI fragment that is 1.65 kb in five of the eight rDNA clones and that lies in the non-coding part of the rDNA repeat unit. We have shown that this fragment is homologous

with restriction fragments in genomic DNA of various lengths. This fragment contains a 600 bp region which is cleaved into small fragments by *Sal*I. In *Xenopus*, most of the variation in rDNA repeat length is due to changes in the number of copies of a short repeat element which acts as an enhancer for RNA polymerase I (Reeder, 1984). The region of the *Arabidopsis* rDNA repeat unit which is cleaved frequently by *Sal*I may contain a similar series of repeat elements. Variations in the number of copies of such an element could account for the heterogeneity of the rDNA repeats.

Most plant rDNAs are heavily methylated at cytosine residues (Gerlach & Bedbrook, 1979; Siegel & Kolacz, 1983; Uchimiya *et al.*, 1982) and the rDNA of another member of the Cruciferae, *Raphanus sativus*, is heavily methylated on the inner cytosine of the CCGG *Hpa*II/*Msp*I recognition sequence, with lesser but significant methylation on the outer cytosine as well (Delseny *et al.*, 1984). The rDNA of *Arabidopsis* is similar to other plants in this respect also, being partially digested by *Hpa*II and to a greater extent by *Msp*I. The methylation experiments with the unique clones demonstrate that there are CCGG sequences which are unmethylated and are cleaved by *Hpa*II and *Msp*I. Furthermore, they show that sites which are cleaved by one isoschizomer but not the other are relatively rare. This contrasts with the situation in wheat where 90% of CCGG sequences are resistant to digestion with *Hpa*II and 50% cannot be cleaved by *Msp*I (Gruenbaum *et al.*, 1981). There may be restriction sites which are methylated on both cytosines and would not be detected by this experiment.

Leutwiler *et al.* (1984) found the *Arabidopsis* genome to contain approximately 14% rapidly reannealing sequences. The most likely types of sequence in this rapidly renaturing component are satellites and foldback sequences. We have not recovered any clones which contain sequences repeated more than 600 times per haploid genome. It is possible that the highly repeated

sequences are present in a sequence organization pattern which prevents them from being cloned in a EcoRI partial digest library. If the quickly reassociating component is due to foldback sequences, they may be present in our clones; we have done no experiments to detect the presence of foldback sequences.

We have recovered four different types of clones which belong to the middle repetitive kinetic component. rDNA clones comprise 16% of the random clones, and the 95% confidence limits for the percentage of such clones present in the entire recombinant library are 6-26%. Chloroplast clones represent 8% (0-16%, 95% confidence interval) and the dispersed repetitive sequences 6% (0-12%, 95% confidence interval). Although the number of clones containing dispersed repetitive sequences constitute 6% of the total, even these clones contain mostly unique sequences. Therefore, the percentage of repetitive sequences is actually less than half of this amount. In addition there is the low copy number repeat found in λbAt119, which represents 2% of the clones examined (0-4%, 95% confidence interval). Thirty-two percent of the clones examined contain repetitive DNA which would reassociate in the middle repetitive component, which compares with 27% of *Arabidopsis* DNA which reannealed in the middle repetitive component in the kinetic analysis of Leutwiler *et al.* (1984). The close correlation of these figures is another indication that the recombinant libraries used in this study are representative of the genomic DNA of the plant. Although the rDNA clones appear to be the major component of the middle repetitive DNA, it is clear from the dot blot data that the total rDNA consists of approximately 6000 kb per haploid genome while the approximately 330 copies of the chloroplast genome must contribute 5-10 times this amount depending on the exact size of the chloroplast genome. The reason for the detection of fewer chloroplast clones than expected is not known. It is possible that the quantity of chloroplast DNA in each DNA preparation varies depending on environmental conditions and the age of the

plants at the time they are harvested.

The remainder of the clones contain DNA sequences which are present only once per haploid genome with the exception of λbAt121 and λbAt131, which appear to contain duplicated sequences. These 34 clones have an average insert length of 13 kb. If we adopt the simple model of sequence interspersion of unique sequences of length x interspersed with repeated sequences of length y, then the frequency of clones of length l containing only unique sequences and coming from the interspersed region is given by:

$$P_{unique} = \frac{x-l}{x+y}.$$

Excluding the rDNA and chloroplast clones we have 34 of 38 clones or 89% which are unique over their entire length. We cannot solve the above equation explicitly because we do not know the length of the average repeat unit. However, if we assume that the repeat unit length is less than or equal to 1 kb, we can calculate a mean unique sequence length of 120-125 kb. If the average repeat unit length is actually longer than 1 kb, we have underestimated the length of the average unique sequence block. Because the number of dispersed repetitive clones is small, the percentage of unique clones is subject to error. If we use the upper limit of the 95% confidence interval for the dispersed repetitive clones (and decrease the percentage of unique clones accordingly), the estimated average unique length would be reduced to approximately 65 kb. Even under these assumptions it is clear that the sequence interspersion pattern of *Arabidopsis* is extremely long. If the average single-copy sequence length is 125 kb, there are fewer than 600 dispersed repeats in the *Arabidopsis* genome. These facts, in combination with the small genome size, make *Arabidopsis* a higher plant uniquely suited to the techniques of molecular genetics. Together with the restriction

fragment length polymorphism genetic map now being constructed it should prove possible to use the method of chromosome walking to clone any genetic locus of interest by starting with a lambda clone that maps near the locus, and successively isolating overlapping clones.

We thank Robin K. Wilson for technical assistance and members of the Meyerowitz laboratory for reading and commenting on the manuscript. This work was supported by grant number 82-CRCR-1-1063 from the Science and Education Administration of the U.S. Department of Agriculture and by grant number PCM-8408504 from the National Science Foundation to E.M.M. R.E.P. was supported by a National Science Foundation pre-doctoral fellowship and by National Research Service Award number 5 T32 GM07616 from the National Institutes of Health.

REFERENCES

Bennett, M. D. & Smith, J. B. (1976). *Proc. R. Soc. Lond. B*, **274**, 227-274.

Davidson, E. H., Hough, B. R., Amenson, C. S. & Britten, R. J. (1973). *J. Mol. Biol.* **77**, 1-23.

Davis, R. W., Botstein, D. & Roth, J. R. (1980). *Advanced Bacterial Genetics*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Delseny, M., Laroche, M. & Penon, P. (1984). *Plant Physiol.* **76**, 627-632.

Denhardt, D. T. (1966). *Biochem. Biophys. Res. Commun.* **23**, 641-646.

Flavell, R. (1980). *Ann. Rev. Plant Physiol.* **31**, 569-596.

Flavell, R. B. & Smith, D. B. (1976). *Heredity* **37**, 231-252.

Gerlach, W. L. & Bedbrook, J. R. (1979). *Nucl. Acids Res.* **7**, 1869-1885.

Graham, D. E., Neufeld, B. R., Davidson, E. H. & Britten, R. J. (1974). *Cell* **1**, 127-137.

Gruenbaum, Y., Naveh-Many, T., Cedar, H. & Razin, A. (1981). *Nature* **292**, 860-862.

Ingle, J., Timmis, J. N. & Sinclair, J. (1975). *Plant Physiol.* **55**, 496-501.

Kranz, A. R. (1978). *Arabid. Inf. Serv.* **15**, 118-139.

Leach, D. R. F. & Stahl, F. W. (1983). *Nature* **305**, 448-451.

Leutwiler, L. S., Hough-Evans, B. R. & Meyerowitz, E. M. (1984). *Mol. Gen. Genet.* **194**, 15-23.

Leweke, B. & Hemleben, V. (1982). *The Cell Nucleus*, Vol. 11 (Busch, H. & Rothblum, L., eds.), pp. 225-253, Academic Press, New York.

Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982). *Molecular Cloning*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Manning, J. E., Schmid, C. W. & Davidson, N. (1975). *Cell* **4**, 141-155.

Meyerowitz, E. M. & Hogness, D. S. (1982). *Cell* **28**, 165-176.

Meyerowitz, E. M. & Martin, C. H. (1984). *J. Mol. Evol.* **20**, 251-264.

Murray, M. G., Cuellar, R. E. & Thompson, W. F. (1978). *Biochemistry* **17,** 5781-5790.

Murray, M. G., Palmer, J. D., Cuellar, R. E. & Thompson, W. F. (1979). *Biochemistry* **18,** 5259-5266.

Peacock, C. & Dingman, C. W. (1968). *Biochemistry* **7,** 668-674.

Redei, G. P. (1965). *Amer. J. Bot.* **52,** 834-841.

Redei, G. P. (1970). *Biblio. Genetica,* **20,** 1-151.

Reeder, R. H. (1984). *Cell* **38,** 349-351.

Rigby, P. W. J., Dieckmann, M., Rhodes, C. & Berg, P. (1977). *J. Mol. Biol.* **113,** 237-251.

Röbbelen, G. (1965). *Arabid. Inf. Serv.* **2,** 36-47.

Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Peterson, G. B. (1982). *J. Mol. Biol.* **162,** 729-773.

Sears, L. M. S. & Lee-Chen, S. (1970). *Can. J. Genet. Cytol.* **12,** 217-223.

Seed, B., Parker, R. C. & Davidson, N. (1982). *Gene* **19,** 201-209.

Siegel, A. & Kolacz, K. (1983). *Plant Physiol.* **72,** 166-171.

Smith, D. B. & Flavell, R. B. (1977). *Biochim. Biophys. Acta* **474,** 82-97.

Uchimiya, H., Kato, H., Ohgawara, T., Harada, H. & Sugiura, M. (1982). *Plant Cell Physiol.* **23,** 1129-1131.

Walbot, V. & Dure, L. S. (1976). *J. Mol. Biol.* **101,** 503-536.

Wyman, A. R., Wolfe, L. B. & Botstein, D. (1985). *Proc. Natl. Acad. Sci. USA* **82,** 2880-2884.

Figure Legends

FIG. 1. Hybridization of *Arabidopsis* genomic DNA to *Arabidopsis* clones. 0.5 μg of DNA from each of the clones λbAt113-118 was digested individually with *Eco*RI and the samples loaded on six adjacent lanes of a 0.8% agarose gel. After electrophoresis the DNA in the gel was denatured and blotted to a nitrocellulose filter. The filter was hybridized with ^{32}P-labeled *Arabidopsis* genomic DNA and autoradiographed. λcI857 S7 DNA digested with *Hin*dIII was used as a molecular weight marker; molecular weights of the lambda fragments are given in kb. A) Ethidium bromide stain of gel. B) Autoradiogram of filter.

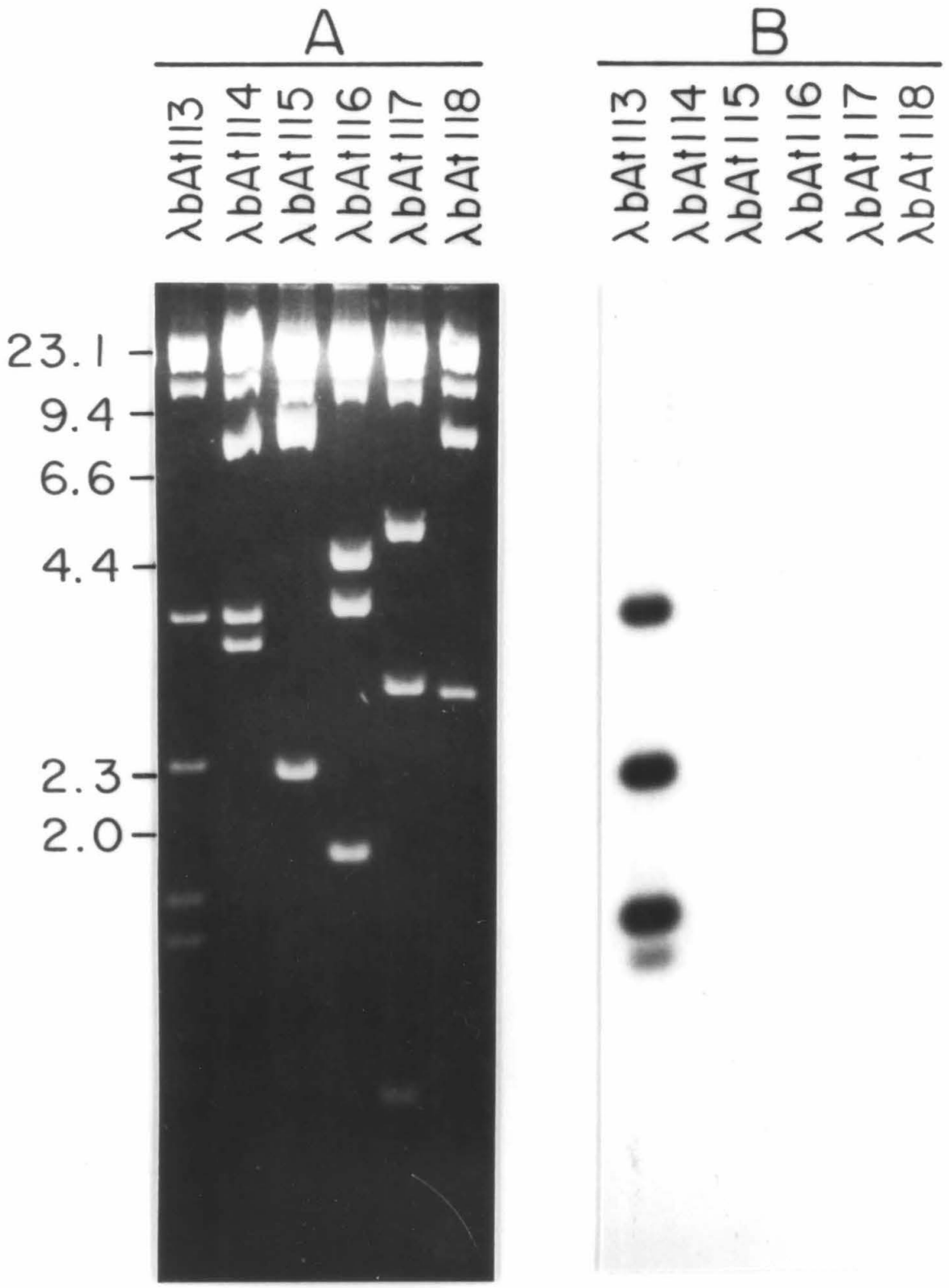

A
B
λbAt113
λbAt114
λbAt115
λbAt116
λbAt117
λbAt118
23.1
9.4
6.6
4.4
2.3
2.0

FIG. 2. Estimation of copy number of *Arabidopsis* genomic clones. DNA samples containing 1.0 µg *Arabidopsis* genomic DNA (lanes labeled G) or 1.0 µg mouse genomic DNA plus DNA from the indicated clone equivalent to 1, 3 or 10 copies per haploid genome (lanes labeled 1, 3 and 10) were digested with *Eco*RI and fractionated on a 0.8% agarose gel. After electrophoresis the DNA in the gel was denatured and transferred to a nitrocellulose filter. The filter was hybridized with ^{32}P-labeled DNA from the same clone which was used to make the filter and autoradiographed. The bands which appear in the lanes containing reconstructions but not in the lane containing the genomic DNA are due to hybridization of the lambda-sequences found in the recombinant clones. λcI857 S7 DNA digested with *Hind*III was used as a size standard; fragment sizes are given in kb.

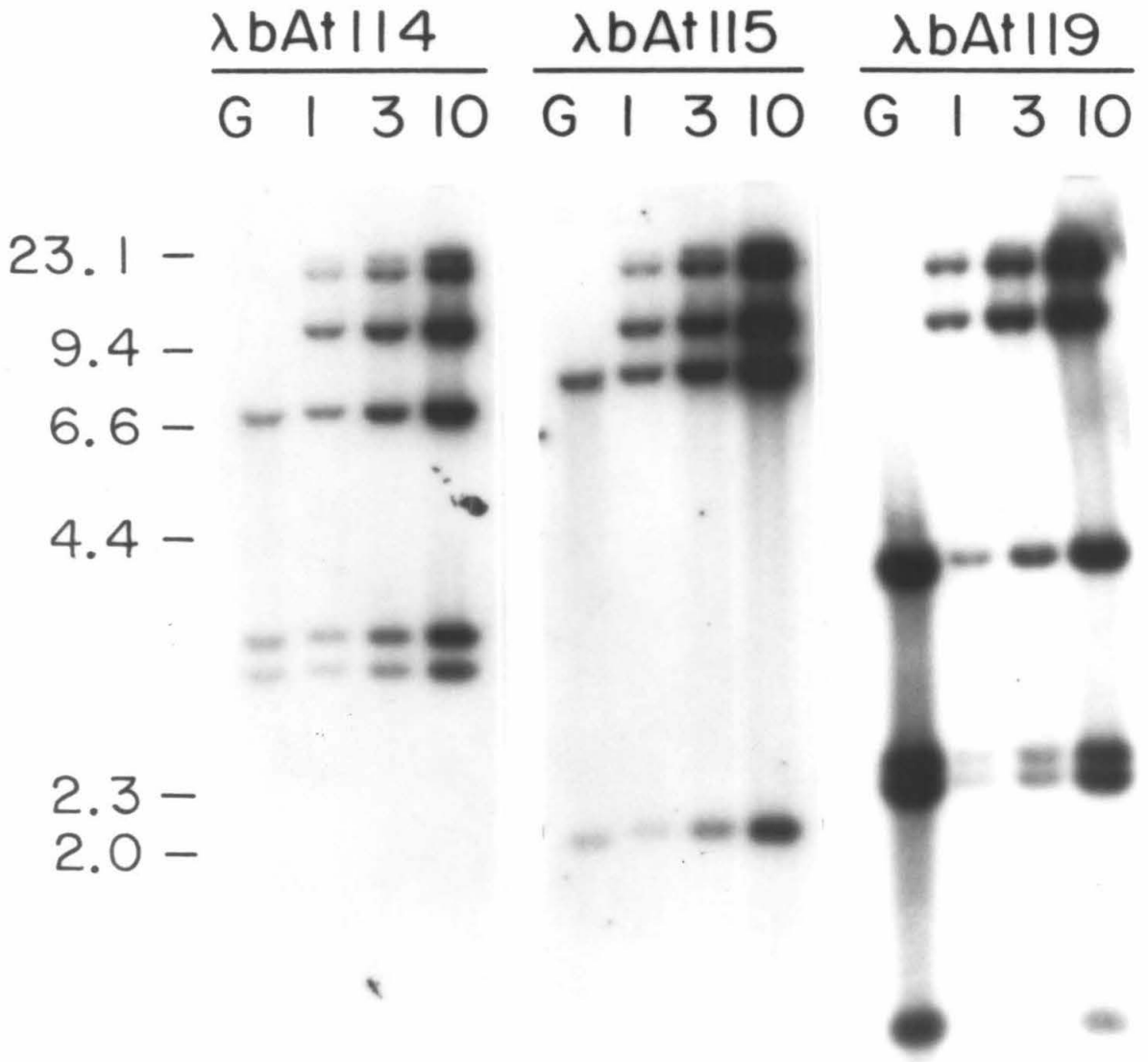

λbAt114
λbAt115
λbAt119
G 1 3 10
G 1 3 10
G 1 3 10
23.1 –
9.4 –
6.6 –
4.4 –
2.3 –
2.0 –

FIG. 3. Restriction map of the *Arabidopsis* rDNA repeat unit. All known restriction sites for *Bgl*II, *Eco*RI, *Xba*I and *Xho*I are shown. The hatched box labeled SR represents a region of the repeat unit which is reduced to fragments shorter than 100 base pairs after digestion with *Sal*I. The rightmost *Eco*RI fragment, which is depicted as being 1.65 kb long in the figure, varies in length in different clones from 1.6 to 2.0 kb.

Eco RI
Eco RI
Xho I
Bgl II
Eco RI
Xba I
Eco RI
SR
Eco RI
1 kb

FIG. 4. Hybridization of the rDNA *Eco*RI restriction fragments to *Arabidopsis* RNA. Total *Arabidopsis* RNA was denatured and fractionated on a 1.5% formaldehyde-agarose gel. The RNA in the gel was base-treated and transferred to a nitrocellulose filter. The filter was cut into five strips and each strip hybridized with ^{32}P-labeled DNA prepared from a different plasmid subclone containing one of the five *Eco*RI fragments of the rDNA repeat unit. After hybridization the filters were autoradiographed.

Arabidopsis 9.9 kb rDNA repeat unit

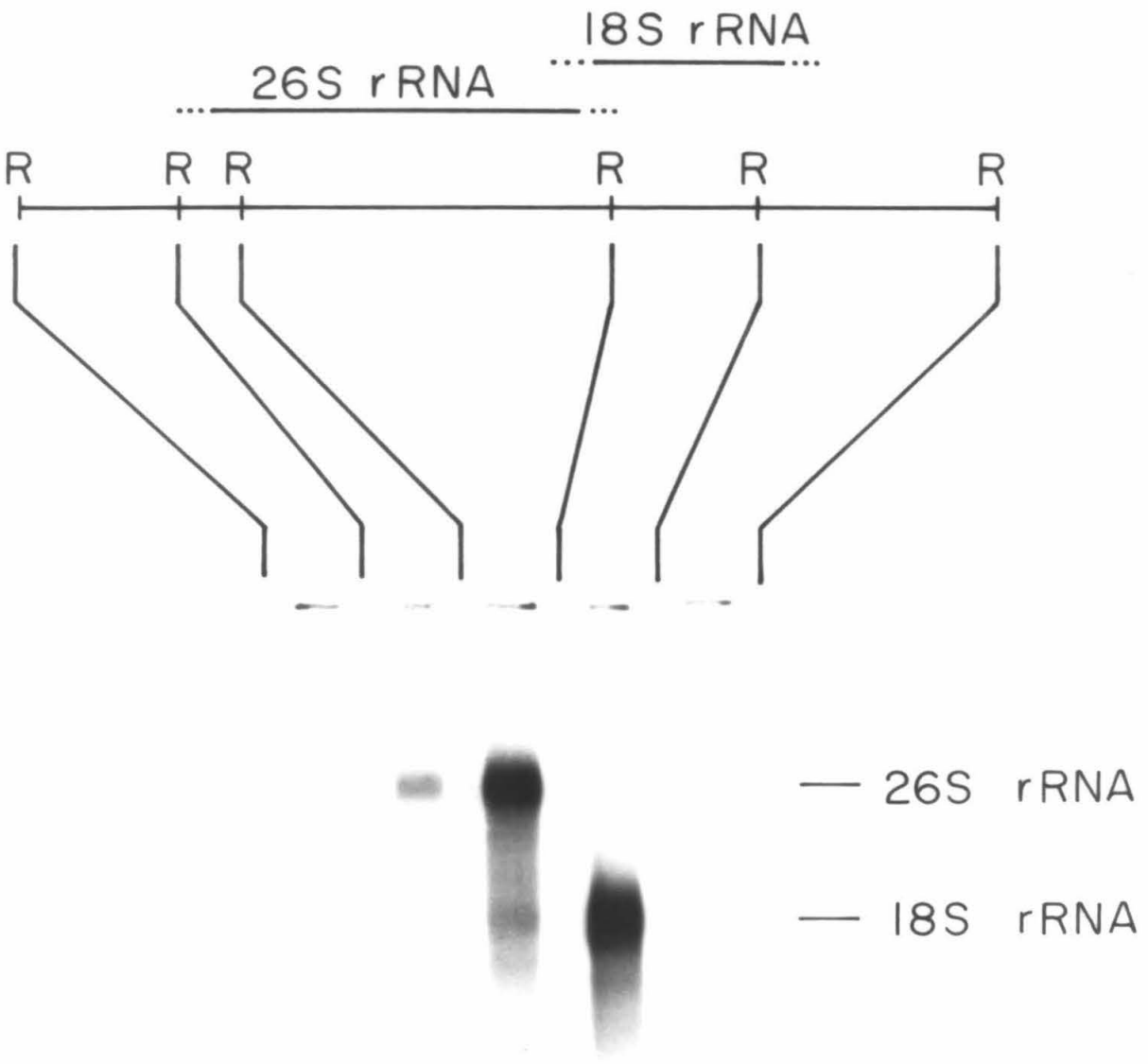


FIG. 5. Quantitation of two classes of *Arabidopsis* repetitive sequences. *Arabidopsis* genomic DNA was serially diluted twofold 12 successive times into a solution of mouse DNA of the same concentration as the starting concentration of the *Arabidopsis* DNA. These 12 dilutions were denatured, neutralized and filtered through nitrocellulose to produce eight identical filters each composed of 12 1.0 μg dots of DNA containing from 1.0 μg to 0.5 ng of *Arabidopsis* DNA. Two filters were hybridized with each of four different ^{32}P-labeled DNA probes: λbAt002, λbAt003, λbAt114 and λbAt115. The concentrations of the probes were identical for each of the four clones. After hybridization the dots were separated and the hybridization quantitated by liquid scintillation spectrometry, as described in Materials and Methods. The data were plotted and lines fitted by linear regression. All points were used except for the points corresponding to the three highest genomic DNA concentrations for λbAt002 and λbAt003, which were non-linear. Points at the lower genomic DNA concentrations are omitted from the figure for clarity. A) λbAt002; B) λbAt003; C) λbAt114; D) λbAt115.

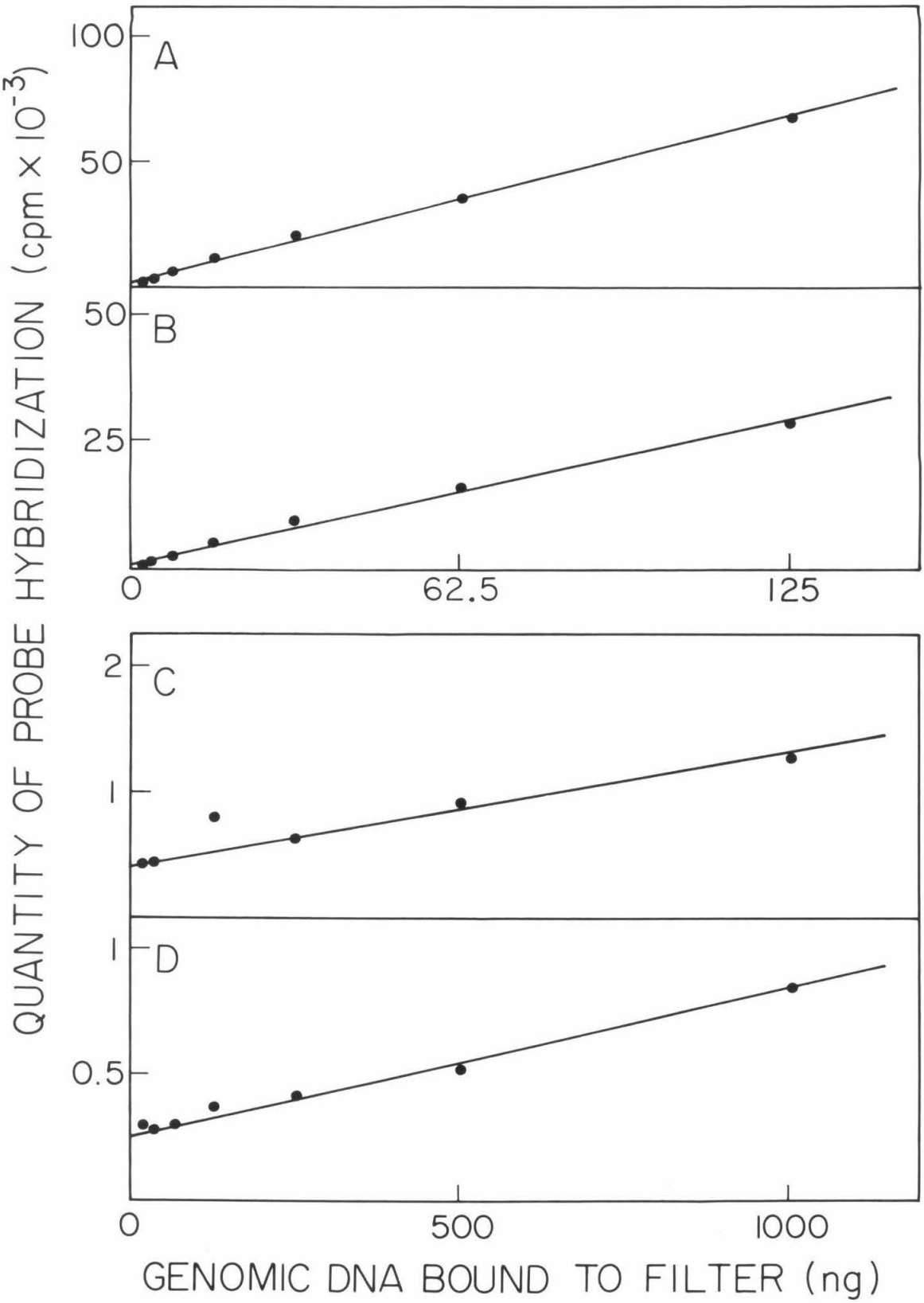

QUANTITY OF PROBE HYBRIDIZATION (cpm x 10⁻³)
100
50
A
50
25
B
0
62.5
125
2
1
C
1
0.5
D
0
500
1000
GENOMIC DNA BOUND TO FILTER (ng)

FIG. 6. Methylation of *Arabidopsis* DNA sequences. 1.0 μg of genomic DNA was digested with either *Hpa*II (lanes marked H) or *Msp*I (lanes marked M) and loaded on two adjacent lanes of a 1.0% agarose gel. After electrophoresis the DNA in the gel was denatured and transferred to a nitrocellulose filter. Four filters were constructed and allowed to hybridize with the ^{32}P-labeled DNA from the clones λbAt103, λbAt104, λbAt105, and λbAt118 and then autoradiographed. Following autoradiography the hybridized DNA was removed from one of these filters and the filter hybridized with ^{32}P-labeled DNA from the clone λbAt025 and autoradiographed. Size markers are from a *Hin*dIII digest of λcI857 S7 DNA; sizes are given in kb.

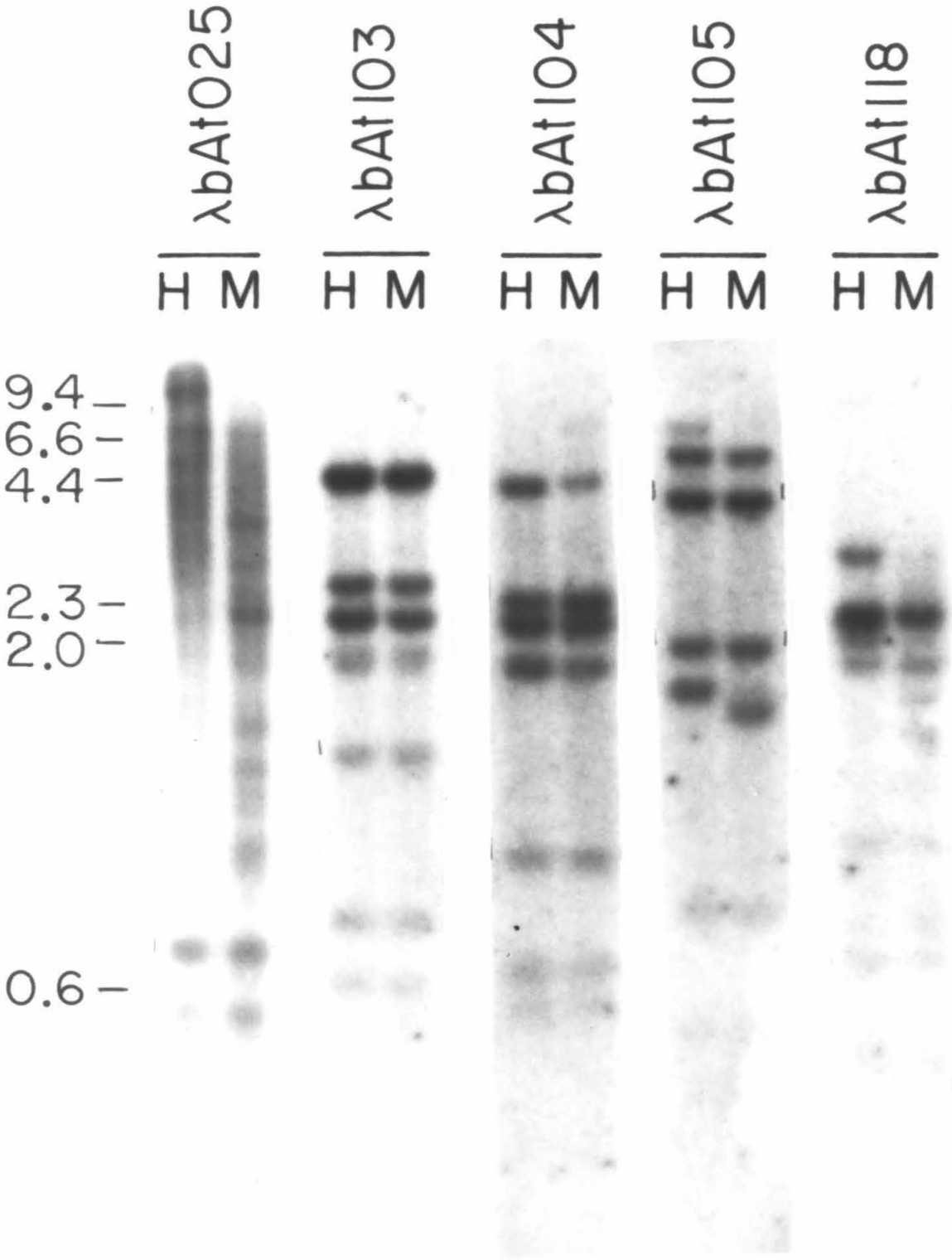

λbAt025
λbAt103
λbAt104
λbAt105
λbAt118
H M
H M
H M
H M
H M
9.4
6.6
4.4
2.3
2.0
0.6

FIG. 7. Identification of chloroplast clones. Approximately 0.5 μg of DNA from the clones λbAt003, λbAt102, λbAt124 and λbAt025 was digested with *Eco*RI and loaded in two sets of four adjacent lanes on a 0.8% agarose gel. After electrophoresis the DNA was denatured and two identical filters were prepared by blotting to nitrocellulose. A) Filter hybridized with ^{32}P-labeled whole plant DNA and autoradiographed. B) Filter hybridized with ^{32}P-labeled purified chloroplast DNA and autoradiographed. The size standards are from λcI857 S7 DNA digested with *Hin*dIII, fragment sizes are given in kb.

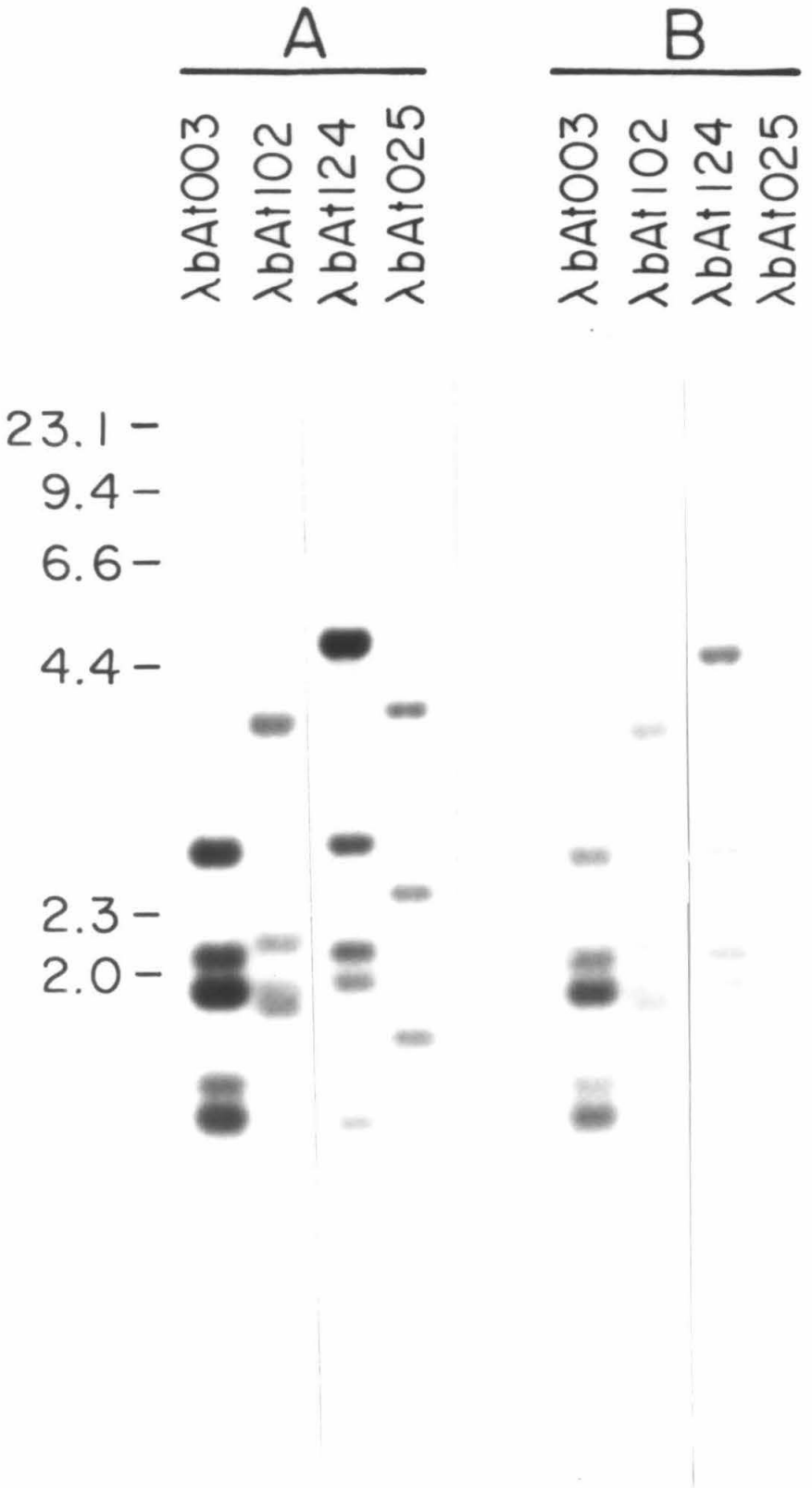

A
B
λbAt003
λbAt102
λbAt124
λbAt025
λbAt003
λbAt102
λbAt124
λbAt025
23.1
9.4
6.6
4.4
2.3
2.0

FIG. 8. Hybridization of λbAt010 and λbAt026 to *Arabidopsis* genomic DNA. Genomic DNA was digested with *Eco*RI and 1.0 μg was loaded in each of two lanes of a 0.8% agarose gel. After electrophoresis the DNA was denatured and each lane was blotted to a separate nitrocellulose filter. One filter was hybridized with ^{32}P-labeled DNA from λbAt010 and the second filter was hybridized with ^{32}P-labeled DNA from λbAt026. The size standards are from λcI857 S7 DNA digested with *Hin*dIII; fragment sizes are shown in kb.

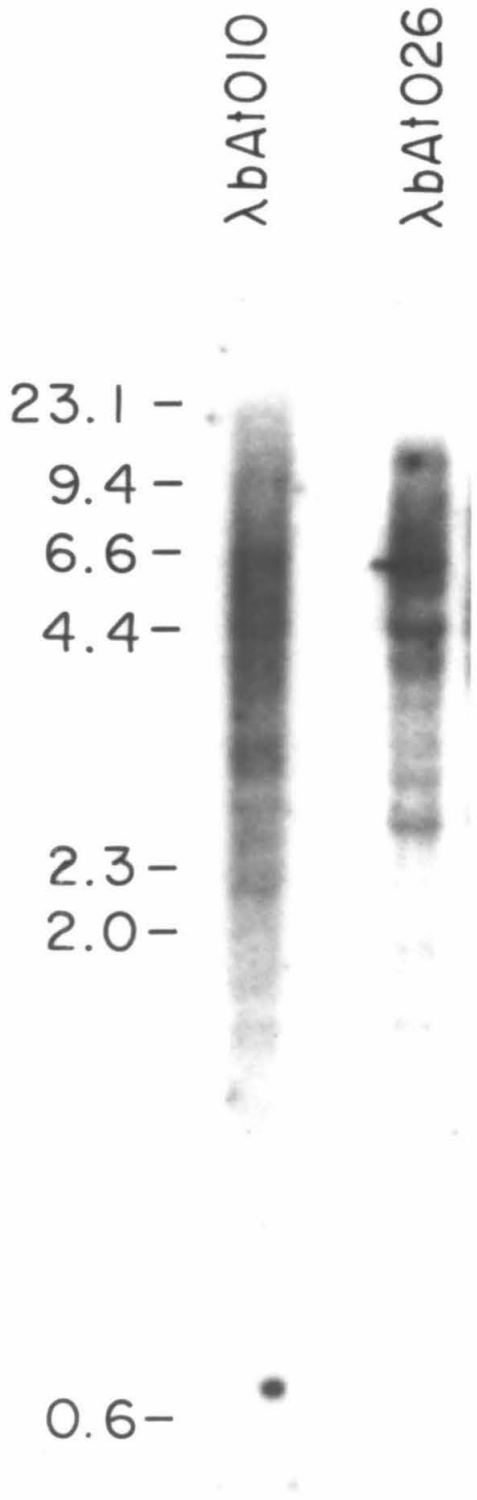

λbAt010
λbAt026
23.1 -
9.4 -
6.6 -
4.4 -
2.3 -
2.0 -
0.6 -

FIG. 9. Restriction maps of λbAt010 and λbAt026. The hatched bars below the maps indicate the regions of these clones which display strong autoradiographic signals when hybridized with ^{32}P-labeled genomic DNA.

λbA1026
Eco RI
Hind III
Eco RI
Hind III
Sal I
Hind III
Bgl II
Sal I
Eco RI
Hind III
Hind III
Eco RI
λbA1010
Eco RI
Xba I
Bam HI
Bam HI
Eco RI
Eco RI
Xba I
Eco RI
Bam HI
Eco RI
Xba I
Xba I
Sal I
Xba I
Xba I
Bam HI
Xba I
Sal I
Xba I
Bam HI
Eco RI
1 kb

Table 1

Classes of Arabidopsis clones (with length of insert in kb)

	Single copy			
Low frequency class	λbAt001 (17.8)	λbAt027 (15.7)	λbAt104 (10.4)	λbAt116 (10.7)
	λbAt009 (16.5)	λbAt033 (12.8)	λbAt105 (11.2)	λbAt117 (9.5)
	λbAt012 (15.9)	λbAt036 (19.9)	λbAt108 (10.3)	λbAt118 (11.7)
	λbAt013 (13.3)	λbAt047 (13.6)	λbAt109 (7.5)	λbAt120 (8.6)
	λbAt014 (16.0)	λbAt049 (8.5)	λbAt110 (15.8)	λbAt126 (13.6)
	λbAt015 (15.6)	λbAt052 (10.7)	λbAt111 (9.7)	λbAt132 (14.0)
	λbAt018 (13.0)	λbAt053 (12.0)	λbAt114 (15.3)	λbAt133 (20.1)
	λbAt020 (11.5)	λbAt103 (12.0)	λbAt115 (11.6)	λbAt134 (9.9)
	Duplicated sequences			
	λbAt121 (9.3)	λbAt131 (11.0)		
	Low copy number repeat			
	λbAt119 (9.1)			

(**Table 1,** continued)

	rDNA			
High frequency class	λbAt002 (14.8)	λbAt106 (11.9)	λbAt113 (13.7)	λbAt123 (10.2)
	λbAt025 (13.7)	λbAt107 (10.1)	λbAt122 (9.9)	λbAt125 (14.8)
	Chloroplast			
	λbAt003 (14.1)	λbAt102 (9.7)	λbAt112 (13.1)	λbAt124 (12.1)
	Dispersed repetitive elements			
	λbAt006 (13.8)	λbAt010 (13.9)	λbAt026 (15.9)	

Chapter 3

Molecular cloning and nucleotide sequence of two genes which are abundantly expressed in the seed of *Arabidopsis*

(draft for submission to Plant Molecular Biology)

Molecular cloning and nucleotide sequence of two genes which are abundantly expressed in the seed of *Arabidopsis*

Robert E. Pruitt and Elliot M. Meyerowitz

Division of Biology

California Institute of Technology

Pasadena, California 91125

Abstract

The experiments described in this paper characterize two pairs of genes from *Arabidopsis thaliana* which are abundantly expressed in the latter half of embryogenesis. One pair of genes encodes the *Arabidopsis* 12S storage protein. The two genes of this pair are arranged as a small tandem duplication in the genome of the Landsberg strain of *Arabidopsis* and are the only two genes in the genome which can be detected by genome blotting experiments. It appears that the Columbia strain of *Arabidopsis* lacks this tandem duplication and presumably contains only a single gene encoding 12S storage protein. These genes are shown to be expressed specifically in seeds during the latter half of embryogenesis. The nucleotide sequence of one copy of the 12S storage protein gene and the 5' end and flanking region of the second gene have been determined and the sequence compared to 12S storage protein genes from *Brassica napus* and *Pisum sativum*.

The second pair of genes characterized encodes a small RNA of only 650 nucleotides the protein product of which is unknown. Once again, the genes are arranged as a small tandem duplication although genome blotting experiments reveal the possible existence of a third gene. These two genes are also shown to be expressed exclusively in seeds during the latter half of seed development although with a slightly different time course than the 12S storage protein genes. The nucleotide sequence which was determined for one of these two genes contains a 148 amino acid open reading frame in the correct orientation relative to the direction of transcription and also contains the typical TATA box upstream of the gene and polyadenylation signal downstream.

Introduction

During the development of the seed of a dicotyledonous plant the embryo develops from a single diploid cell surrounded by endosperm into a small plantlet composed of two "seed leaves" or cotyledons and the rudiments of both shoot and root. During this process the embryo increases in size and also differentiates specialized tissues: the maturing embryo contains distinct epidermal and vascular tissues as well as the obviously differentiated root and shoot apices. At the end of embryogenesis the entire seed undergoes desiccation allowing the embryo to withstand a wide variety of conditions for an extended period without loss of viability.

In the latter half of embryogenesis the most abundant messenger RNAs are those encoding the seed storage proteins. The storage proteins accumulate to high levels in the seed and are rapidly broken down when the seed germinates. They are thought to serve as a store of nitrogen for the developing seedling. Most plants produce several different storage proteins in their seeds, each of which may be expressed with its own particular temporal pattern and many of which are encoded by multigene families. Because the storage proteins constitute a major source of edible protein and are generally deficient in one or more essential amino acids it would be desirable to be able to modify the genes encoding these proteins. Study of homologous proteins from various plant species may allow the identification of regions within the storage protein which may be modified without interfering with any of the structural or functional requirements of the protein.

Many dicotyledonous plant seeds contain a 12S storage protein. These proteins consist of two polypeptides, designated α and β, which, like most storage proteins, are located in membrane bound protein bodies in the embryo (Pernollet, 1978). Both polypeptides are encoded by the same gene and are produced by

posttranslational cleavage of a common precursor protein (Higgins, 1984). This precursor also contains a hydrophobic amino terminal leader peptide which is not present in the mature proteins that is presumably responsible for localizing the storage proteins in the protein bodies. The 12S storage protein from pea (*Pisum sativum*) has been shown to be homologous with that from oilseed rape (*Brassica napus*) and, although the amino acid sequences are quite divergent, there are clearly regions which are well conserved as well as regions in which large insertions are present in one protein relative to the other (Simon et al., 1985). Both of these genes are encoded by multiple genes: there are thought to be 4 genes encoding the pea 12S storage protein (Croy et al., 1982) and either 3 or 4 encoding the *B. napus* 12S protein (M.L. Crouch, personal communication).

In the experiments described in this paper we have cloned at least five different classes of abundantly expressed seed RNA species from *Arabidopsis thaliana* . Two of these classes were characterized with respect to the organization of the genome in the region containing the genes, the time and tissues of expression and finally with respect to nucleotide sequence. One of these classes represents the *Arabidopsis* 12S storage protein genes while the other class codes for an unknown product.

Materials and Methods

Plant culture

The strains of *Arabidopsis* used in the experiments described in this paper were: Landsberg ***erecta***, obtained from F. J. Braaksma, Department of Genetics, Biology Centre, Haren, The Netherlands; and Columbia, obtained from A. Kleinhofs, Program in Genetics, Washington State University, Pullman, WA

99164. The Landsberg strain used in these experiments bears a homozygous recessive mutation, *erecta*, which makes the plants more compact and easier to culture in large numbers.

Plants were cultured as described in Pruitt and Meyerowitz (1986). Plants which were used for determining the time of expression during seed development were grown at a density of five plants per 2.25 inch pot at 23.5-24.5^{o}C, 70% relative humidity and 7000 lux constant illumination.

General DNA and recombinant DNA techniques

Restriction digestions were performed as described by Davis *et al.* (1980). ^{32}P-labeled DNA was prepared by nick translation using the method described by Rigby *et al.* (1977). PolyA^{+} RNA and ^{32}P-labeled cDNA were prepared as described by Meyerowitz and Martin (1984). Recombinant lambda libraries were screened as described in Meyerowitz and Martin (1984). Preparation of bacteriophage and plasmid DNA was by the method of Davis *et al.* (1980). *Arabidopsis* DNA and RNA were prepared by the method we have previously described (Pruitt and Meyerowitz, 1986). *In vitro* transcriptions using Sp6 or T7 RNA polymerase were performed as recommended by the manufacturer (Promega Biotech).

Electrophoresis, filter binding and hybridization of nucleic acids

Electrophoresis, filter binding and hybridization of nucleic acids were performed as described in Meyerowitz and Martin (1984) except for hybridization of single stranded RNA probes to RNA filters. Hybridizations between RNA probes and RNA filters were performed in 50% (v/v) formamide, 5xSSPE (1xSSPE is 180 mM NaCl, 10 mM NaH_2PO_4, 8 mM NaOH, 1 mM Na_2EDTA, pH 7.0; Davis *et al.*, 1980), 100 µg/ml sonicated and denatured salmon testis DNA, 500 µg/ml yeast RNA, 1xDenhardt's solution (0.02% ficoll, 0.02% polyvinylpyrrolidone, 0.02% bovine serum albumin; Denhardt, 1966), and 0.1% SDS at 60°C. After hybridization these filters were washed in 0.05xSSPE and 0.1% SDS at 55°C.

Nucleotide sequencing

Nucleotide sequencing was performed using the chemical method of Maxam and Gilbert (1980) essentially as described in Garfinkel et al. (1983). Labeling of ends which have a 3' protruding end was done either with terminal transferase and α-^{32}P-dideoxyATP (for 3' ends) as described by the commercial supplier (Amersham) or using T4 polynucleotide kinase as described below.

Plasmid DNA (10 µg) was digested with a twofold excess of restriction enzyme in a volume of 50 µl. At the end of the incubation 1.2 units of T4 DNA polymerase were added to the reaction and the incubation continued for an additional 90 seconds. The reaction tube was then immediately moved to a preheated 70°C heat block and incubated for 10 minutes to denature the T4 DNA polymerase. The resulting protruding 5' end was then treated with 0.5 units of calf intestinal alkaline phosphatase for 30 minutes followed by addition of 5 µl of 0.5 M EDTA. The reaction was then phenol-extracted, chloroform-extracted and ethanol-precipitated. The resulting DNA was labeled with T4 polynucleotide kinase, digested with a second restriction enzyme and used for DNA sequencing.

Results

Cloning of genes encoding abundant seed RNA species

In order to isolate molecular clones of genes which are abundantly expressed in the developing *Arabidopsis* seed, ^{32}P-labeled cDNA was prepared from *Arabidopsis* seed pods and used as a hybridization probe to screen a recombinant lambda library containing *Arabidopsis* genomic DNA. Four to six genome equivalents of recombinant phage were plated and nitrocellulose filter replicas made of the plates. PolyA$^+$ RNA was isolated from 7-8 day old *Arabidopsis* seed pods and used to prepare ^{32}P-labeled cDNA which was allowed to hybridize to the filter replicas. After hybridization the filters were washed and autoradiographed. 18 plaques which showed a positive hybridization signal were picked from the phage plates and the recombinant phage eluted from the agar plugs. These phage were replated and duplicate nitrocellulose filters were prepared from each plate. ^{32}P-labeled cDNA was again used as a hybridization probe on one set of filters and ^{32}P-labeled DNA from a cDNA clone containing sequences from the *Brassica napus* 12S storage protein gene (pC1; Simon *et al.*, 1985) was used to screen the other set of filters. Of the 18 original positive clones 14 phage gave positive signals when rescreened with labeled cDNA and 3 recombinant phage (λfAt1504, λfAt1505 and λfAt1515) also hybridized to the *B. napus* storage protein gene probe. It is not known if the four phage which failed to give positive signals on a second screening were falsely positive on the first screening or if they were genuine positives which were lost when the original plaques were picked and the phage replated.

Preliminary restriction digestion and cross-hybridization experiments

indicated that the 14 clones recovered fall into at least five classes based on sequence homology. Two of these classes, one containing the clones homologous with the *B. napus* 12S storage protein cDNA clone (group I) and the other class selected arbitrarily (group II), were chosen for further analysis. The remaining classes of clones will not be considered further in this paper.

Characterization of the genomic region containing the genes for the Arabidopsis 12S storage protein

In order to facilitate further characterization and sequencing of the genes contained in the clones homologous with the *Brassica napus* 12S storage protein cDNA clone, the recombinant phage were restriction-mapped and the regions homologous with the *B. napus* gene located by blotting experiments. The restriction fragments found in these three clones (λfAt1504, λfAt1505 and λfAt1515) were found to be identical. Since the library from which these clones were isolated was prepared by MboI partial digestion and was amplified before use, these three clones probably represent separate isolates of the same phage. The restriction map of this phage overlaps the restriction map of two recombinant phage (λbAt1501 and λbAt1502) which we had previously isolated from a different library, made by EcoRI partial digestion of *Arabidopsis* DNA from the strain Columbia, using the *B. napus* cDNA clone pC1 as a probe (R.E. Pruitt, unpublished experiments). The restriction maps of the region surrounding the 12S storage protein genes from these two strains is presented in Figure 1. The structure of the restriction maps suggests that there is a tandem duplication present in the Landsberg genome which is absent from the genome of the Columbia strain.

DNA from the phage λbAt1501 was digested separately with EcoRI and XhoI, fractionated on an agarose gel and transferred to a nitrocellulose filter.

This filter was probed with ^{32}P-labeled pC1 DNA, washed and autoradiographed. The autoradiogram demonstrated hybridization of the probe to a 1.6 kb XhoI fragment as well as to two EcoRI fragments of 2.2 and 2.4 kb. The region of hybridization is shown in Figure 1. DNA from the phage λfAt1505 was digested with BglII, EcoRI and HindIII, fractionated by agarose gel electrophoresis and transferred to a nitrocellulose filter. This filter was hybridized with ^{32}P-labeled DNA from sAt2105, a plasmid subclone of λbAt1501 containing the 1.6 kb XhoI fragment which hybridizes to the *B. napus* 12S storage protein cDNA clone. Autoradiography of this filter revealed hybridization of this probe to BglII fragments of 4.3 and 5.0 kb, EcoRI fragments of 1.1, 1.25, 2.4 and 3.2 kb, and HindIII fragments of 2.8, 4.3 and 6.5 kb. As Figure 1 illustrates this pattern of hybridization suggests that the 12S storage protein genes are located in the region which is apparently duplicated in the Landsberg ecotype.

To see if these represented the only genes for this storage protein in the *Arabidopsis* genome, a genomic blot with this same subclone as a probe was performed. *Arabidopsis* genomic DNA from the Landsberg strain was digested with BglII, EcoRI and HindIII, separated by agarose gel electrophoresis and blotted to nitrocellulose. This filter was allowed to hybridize with ^{32}P-labeled sAt2105 plasmid DNA, washed and autoradiographed. The autoradiograms demonstrated hybridization to BglII restriction fragments of 3.8 and 4.3 kb, EcoRI fragments of 1.1, 2.2, 2.4 and 3.2 kb, and HindIII restriction fragments of 2.8, 4.3 and 6.1 kb. All of these fragments correspond to those located in λfAt1505 with the exception of those which are located on the boundary of the clone and therefore would not be expected to correspond with the size of those fragments contained in the genome. This indicates that there are two genes for this storage protein in the genome of the Landsberg ecotype of *Arabidopsis* and that the region of the genome containing them is faithfully represented in the DNA we have cloned.

To determine the direction of transcription of the genes, single-stranded RNA probes were hybridized to RNA blots. *Arabidopsis* seed pod RNA was isolated, denatured and fractionated on a formaldehyde-agarose gel. Following electrophoresis the RNA was transferred to a nitrocellulose filter, the filter cut into two identical strips and the filters allowed to hybridize with ^{32}P-labeled RNA transcribed by Sp6 RNA polymerase from either sAt2101 or sAt2105. These two clones contain the identical 1.6 kb XhoI fragment in opposite orientations downstream of an Sp6 polymerase promoter. After the filters were washed and autoradiographed the filter hybridized with RNA prepared from sAt2105 showed a single band of hybridization corresponding to an RNA of approximately 1700 nucleotides, while no hybridization was apparent on the filter hybridized with RNA from sAt2101 (figure 2). This indicates that the genes are transcribed in the directions indicated in Figure 1.

Characterization of the genomic region contained in the group II clones

A clone representing group II (λfAt1506) was also restriction-mapped and the coding sequences mapped using ^{32}P-labeled cDNA. The restriction map of the region surrounding the cDNA homologous region is presented in Figure 3. DNA from the clone was digested with BglII, EcoRI and HindIII, fractionated by agarose gel electrophoresis and transferred to nitrocellulose. When this filter was allowed to hybridize to ^{32}P-labeled cDNA prepared from seed pods, bands representing two fragments were seen in each of the digests. The fact that hybridization to two fragments was detected in every digest suggested that there might be two genes in this region also. To test this hypothesis the labeled cDNA was thermally removed from the filter and the filter rehybridized with ^{32}P-labeled DNA from the plasmid nAt1510 which contains a 1.9 kb HindIII fragment which hybridized to

the cDNA (see Figure 3). This probe hybridized to the same set of fragments as the labeled cDNA and in every digest hybridized to a greater degree to one of the fragments than to the other. The fragments which hybridize more intensely represent the DNA which was known to be homologous with this fragment based on the restriction map. The other fragments represent the other region which hybridizes with the labeled cDNA. This experiment indicated that these two regions either represent distinct homologous genes or represent separated parts of one gene which contains internally repetitive sequences.

To further characterize these two small regions which hybridize to ^{32}P-labeled cDNA prepared from seed pods, fragments containing each of the regions were subcloned and restriction-mapped in greater detail (Figure 4). These restriction maps demonstrate that the two regions contain a similar restriction map segment of approximately 300 nucleotides, which represents a direct repeat in the genomic DNA cloned. DNA from these plasmid clones (nAt1510 and nAt1511) was digested with various enzymes, fractionated by agarose gel electrophoresis and transferred to nitrocellulose. This filter was allowed to hybridize with ^{32}P-labeled cDNA prepared from seed pods, washed and autoradiographed. The autoradiograms of this blot indicate that the DNA sequences contained in the region of the shared restriction map do in fact hybridize to the labeled cDNA as well as does some sequence extending beyond the SacI site. The extent of cDNA hybridization is indicated in Figure 4.

To determine if this clone also represents the structure of the DNA found in the *Arabidopsis* genome and to determine the copy number of this gene in the genome, a genomic DNA blotting experiment was performed. *Arabidopsis* genomic DNA was digested with BglII, EcoRI and HindIII, subjected to agarose gel electrophoresis and transferred to a nitrocellulose filter. This filter was hybridized with ^{32}P-labeled DNA from the plasmid nAt1510, washed and

autoradiographed. The autoradiogram of this filter revealed hybridization to the restriction fragments expected from the region surrounding nAt1510 but also additional fragments which do not correspond with those from the region surrounding nAt1511. A longer exposure of the filter demonstrated that the restriction fragments of the sizes expected from the nAt1511 region can be detected but at a much lower signal intensity than expected. The exact reason for this is not clear but it is possible that the DNA contained in λfAt1506 represents a minor variant in the *Arabidopsis* genome with respect to sequence organization of this region and the unaccounted for restriction fragments which hybridize more strongly with the probe represent the most common sequemce organization in this region.

To determine the direction of transcription of the gene or genes in this region an RNA blotting experiment was performed using single-stranded probes, similar to that described above for the 12S storage protein gene. *Arabidopsis* seed pod RNA was denatured, fractionated on a formaldehyde-agarose gel and transferred to a nitrocellulose filter. The filter was cut into two strips and each strip was hybridized with a different ^{32}P-labeled RNA probe. The probes were prepared by transcription using either Sp6 polymerase or T7 polymerase from the clone At1519, a plasmid containing the central SacI-SphI fragment from the plasmid nAt1510 located between the Sp6 and T7 RNA polymerase promoters of the vector pGEM-3. Autoradiography of the two filter strips revealed that the probe prepared with Sp6 polymerase hybridized to an RNA species of approximately 650 nucleotides while the probe prepared with the T7 polymerase failed to hybridize to the filter (Figure 2). This indicates that the two genes are transcribed in the directions indicated in Figure 3. The shared restriction map together with the size of the region in each clone which hybridizes to labeled cDNA makes it likely that there are two separate genes which encode the 650

nucleotide RNA detected on the RNA blots.

Tissue specific transcription of both pairs of genes

To determine more precisely the tissue of expression of these genes an RNA blot experiment was performed. RNA was prepared from three different plant materials: seeds, the pod tissue surrounding the seeds and from vegetatively growing plants which contain leaves, stems and roots. The RNA samples were denatured and fractionated on a formaldehyde-agarose gel. After electrophoresis the RNA was transferred to a nitrocellulose filter which was allowed to hybridize with ^{32}P-labeled cloned DNA from a plasmid specific for either the 12S storage protein genes (nAt1512) or the group II genes (nAt1510). In both cases hybridization was observed to a single band of the appropriate size in the RNA prepared from seeds, but no hybridization was observed to RNA made from either the pod tissues or the vegetative plant tissues (Figures 5a, 5b). To be certain that there was in fact RNA present on the filter in the lanes showing no hybridization, the probe was removed and ^{32}P-labeled DNA from a λ clone containing the *Arabidopsis* rDNA was allowed to hybridize to the filter (λbAt106; Pruitt and Meyerowitz, 1986). This probe hybridizes in all three lanes (Figure 5c) thus indicating that there is RNA present in the lanes that show no hybridization in the previous experiment.

Time of expression during the development of the seed

A second RNA blot experiment was performed to determine when in the course of development of the *Arabidopsis* seed each of these two types of genes is expressed. RNA was prepared from seed pods of various ages as well as from

unfertilized buds. These RNA samples were denatured and subjected to electrophoresis through a formaldehyde-agarose gel followed by blotting to a nitrocellulose filter. This filter was probed first with a probe specific for the group II genes (nAt1510) and following removal of that probe with a probe specific for the 12S storage protein genes (nAt1512). After removal of the 12S storage protein gene probe a probe for the rRNA (λbAt106) was used to verify that there was RNA in every sample loaded on the gel. Individual embryos were dissected from pods of the same ages as those used to make the RNA samples to verify the rate at which the embryos were developing. These dissections indicated that the rate of development of the embryos for the Landsberg ***erecta*** strain grown under these conditions is approximately the same as that for the Columbia strain described by Meinke and Sussex (1979). The seeds develop over a period of roughly two weeks with the seeds being mature and desiccated in 14 or 15 days.

Figure 6a shows the results of the experiment using nAt1510 as a hybridization probe. RNA homologous with this probe is detected in the lanes representing 7-8, 9-10 and 11-12 day old embryos. No homologous RNA is detected either before or after the indicated time frame. Figure 6c shows the same filter probed with the rDNA probe revealing that there is RNA present in every lane. However, little can be inferred about the quantitative levels of the RNA since the RNA could be specific to a single tissue within the seed which is changing in size during the course of embryonic development. The rRNA control indicates only that the RNA samples used did contain RNA which was prepared from total seed pods of various ages.

Figure 6b shows the result obtained when this same filter was allowed to hybridize with ^{32}P-labeled DNA from the plasmid clone nAt1512 which contains sequences homologous to the ***Arabidopsis*** 12S storage protein genes. In this case RNA homologous with the probe sequences can be detected in seed pods which are

7-8 days old or older. The seeds contained in the 15-16 day old seed pods are already substantially mature, and therefore this experiment demonstrates that RNA transcribed from the 12S storage protein genes continues to be present in measurable quantities at the end of embryogenesis when the seed is undergoing desiccation. Once again, it is not appropriate to draw quantitative conclusions from this experiment. However, it is clear that the levels of accumulated RNA for the two different pairs of genes follow different temporal patterns during embryogenesis.

Nucleotide sequence of a region containing a group II gene

In order to further characterize the nature of the small genes the nucleotide sequence of one of the regions containing a small gene was determined. The region which contains sequences homologous to cDNA prepared from seeds in the plasmid nAt1511 was sequenced by the chemical method of Maxam and Gilbert (1980). The sequence was determined on both complementary strands and is presented in Figure 7. The sequence contains a TATA box beginning at position 109 and the consensus polyadenylation signal AATAAA at position 769. Assuming that transcription begins approximately 25 nucleotides downstream of the TATA sequence and the mature message terminates within 20-40 nucleotides of the AATAAA, as is the case for most eukaryotic genes, the mature mRNA would be approximately 660 nucleotides long. This is in good agreement with the estimated RNA size of 650 nucleotides. The putative transcript contains a 148 amino acid open reading frame beginning with a methionine codon 120 nucleotides downstream of the TATA sequence. The translation product encoded by this reading frame would have a molecular weight of 14800 daltons. However, the translation product begins with a very hydrophobic stretch of 21 amino acids

which may constitute a leader peptide and therefore might not be included in the mature protein. The remainder of the protein would also be quite hydrophobic with the exception of the carboxy terminus which contains a very high percentage of basic residues. A search of both nucleic acid and protein sequence databases failed to disclose any known sequences homologous with this gene.

Between the probable site of initiation of transcription and the beginning of the open reading frame there exists a second ATG codon at which translation might initiate. If translation were to initiate at that site, it would terminate after producing a peptide of only 18 amino acids. Termination would occur at a position 26 nucleotides upstream of the ATG which begins the long open reading frame. Similar situations are known to exist in a mouse gene (Kahana and Nathans, 1985) and in particular in yeast, where they are thought to be involved in translational control of messenger RNAs (Thireos *et al.*, 1984; Hinnebusch, 1984).

Nucleotide sequence of an Arabidopsis 12S storage protein gene

The complete nucleotide sequence of one of the two *Arabidopsis* 12S storage protein genes was determined by the chemical method of Maxam and Gilbert (1980) and is presented in Figure 8. The sequence was determined on both complementary strands with the exception of 400 nucleotides from the EcoRI site extending toward the 5' end of the gene which was determined only on one strand. By comparison of the *Arabidopsis* gene sequence with the sequence of a cDNA clone encoding *Brassica napus* 12S storage protein (Simon *et al.*, 1985) it is possible to locate the translation initiation codon, the translation termination codon and also the locations of three probable intron sequences. Each of these three intron sequences represent an insertion in the *Arabidopsis* sequence relative to the *B. napus* cDNA sequence and each insertion begins with the nucleotides GT

and ends with the nucleotides AG. In addition, each of these regions contains termination codons in frame with the translated reading frame and would result in premature termination of the protein if they were not processed out of the mature RNA. Comparison of the positions of these introns with the sequence of a genomic clone for *Pisum sativum* 12S storage protein (legumin; Lycett *et al.*, 1984) indicates that they are in precisely the same positions relative to the protein as the introns in the legumin gene. Each intron in the *Arabidopsis* gene is slightly longer than its counterpart in the legumin gene although all the introns are shorter than 130 nucleotides. The region of sequence upstream of the initiating ATG contains a perfect TATA box beginning at position 61 and the region downstream of the termination codon contains a polyadenylation signal beginning at position 2015.

The deduced protein sequence of the *Arabidopsis* 12S storage protein is given in Figure 9, which shows the alignment of this protein with the 12S storage proteins from *Brassica napus* and *Pisum sativum*. The *Arabidopsis* protein sequence is much more similar to the *B. napus* sequence as would be expected from the phylogenetic relationships of these three species. When the proteins from *B. napus* and *P. sativum* are compared, one major difference is the presence of a large insertion in the coding sequence of each gene relative to the other (Simon *et al.*, 1985). The insertion in the *B. napus* gene is characterized by a high percentage of the amino acids glutamine and glycine and is located following position 122 in the *B. napus* protein sequence. The insertion in the *P. sativum* sequence is characterized by an abundance of the amino acids glutamic and aspartic acid and is located between residues 296 and 297 in the *Brassica napus* 12S storage protein sequence. As can be seen in Figure 9 the *Arabidopsis* gene has insertions in both of these locations although neither of the insertions is as large as the insertions in the genes from the other species. The amino acids which

characterize the insertions in the other proteins are also present in the small insertions found in the *Arabidopsis* protein sequence. The *Arabidopsis* protein is actually the smallest of the three proteins although it does not appear to contain the minimum amount of amino acid sequence at either of these two locations. The *Arabidopsis* protein also has an insertion of six amino acids at the junction of the leader peptide and the amino terminus of the mature peptide. A diagrammatic representation of the three genes depicting the positions and relative sizes of the major insertions and exon-intron structure is given in Figure 10.

The nucleotide sequence of the 5' flanking region and 5' end of the second gene was also determined and is also presented in Figure 8. As can be seen there is a varying degree of homology over the entire length of the region sequenced in both genes. The region extending downstream from the ATG which serves as the initiator of translation is conserved perfectly nucleotide-for-nucleotide except for the sequence of the first intron. Immediately upstream of the initiation codon there is evidence of at least one insertion event which presumably lies within the 5' untranslated region of the RNA. Farther upstream there is good sequence homology between the two genes, especially in the region surrounding the TATA box and the region immediately downstream where transcription initiation presumably occurs. Upstream of this region the degree of homology varies but never disappears entirely over the extent of the region sequenced. This is in sharp contrast to the 5' flanking regions of three *Pisum sativum* legumin genes which have been sequenced and show absolute sequence homology for more than 300 nucleotides upstream of the intiation of translation (Lycett *et al.*, 1985).

Discussion

The experiments described characterize two small gene families which are expressed in the developing seed of *Arabidopsis thaliana*. The family of genes which we have called group II encodes a small 650 nucleotide RNA species and is composed of two clustered genes arranged in tandem and possibly a third gene which is not located within the boundaries of the region we have cloned. From the genome blotting experiments which were performed it is also possible, and perhaps more likely, that the extra bands of hybridization detected on the genome blot filters are due to restriction fragment length polymorphisms in one of the two genes which we have cloned. The lack of common restriction sites except in the region containing the transcribed sequences suggests that this gene duplication is not recent. It is not known whether both of these genes are expressed. RNA homologous with these genes is detected only in seeds, not in the surrounding seed pod or in vegetative plant tissues. This RNA is present at levels detectable by RNA blotting experiments at approximately days 7 through 12 of *Arabidopsis* embryogenesis. No RNA is detectable in 5-6 day old seeds or in seeds older than 12 days.

The second pair of genes encodes the *Arabidopsis* 12S storage protein. These two genes are also located in a small cluster in the Landsberg strain of *Arabidopsis* and they are the only genes encoding the 12S storage protein of *Arabidopsis*. In the other plant species from which 12S storage protein genes have been cloned, genomic lambda clones contain only a single gene, indicating that this type of close clustering is not typical of this gene family (Fischer and Goldberg, 1982; Lycett *et al.*, 1984; M. L. Crouch, personal communication). These *Arabidopsis* genes represent a tandem duplication of approximately 4.3 kb, which contains a number of conserved restriction sites as well as some sites

present in only one copy of the duplicated sequences or the other. The presence of conserved restriction sites outside the transcribed region suggests that this gene duplication is more recent than the duplication of the group II genes described above. The presence of sites that are unique to one copy of the sequences is an indication that the copies have diverged since the duplication event. The sequence of the 5' ends of the two genes upstream of the initiation codons is not highly conserved, confirming that the duplication event is not very recent. The sequences surrounding the TATA box and immediately downstream from it, where transcription initiation presumably occurs, are very highly conserved between the two gene copies. This may indicate the importance of these sequences for expression of the genes, although it must be remembered that it has not been demonstrated that both of these genes are expressed. The sequence downstream of the initiating ATG is conserved perfectly over the region sequenced except for the sequence of the first intron which differs both by substitutions and small deletions.

The clones which we have obtained from a second strain of *Arabidopsis*, Columbia, are interesting because they appear to lack the tandem duplication present in Landsberg. This finding is based on the restriction maps of the clones and hybridization of a *B. napus* 12S storage protein cDNA clone to the DNA present in these clones. It is particularly interesting because, if one examines the restriction sites which are unique to one repeat or the other in the Landsberg genome, the sites which are unique to the 5' end of the right gene and the sites unique to the 3' end of the left gene are present in the single gene found in the Columbia strain. This may indicate that the tandem duplication present in the Landsberg strain has been eliminated in the Columbia strain recently, possibly by an unequal crossing-over event. The Columbia strain was derived from a plant of the Landsberg strain by Redei (1970) but it is not known if the Landsberg strain

can be traced to a single plant and therefore it is not known how long these two strains have been diverging. If the Columbia strain does contain only a single gene for the 12S storage protein, it represents a much simpler experimental system to manipulate than the multigene families commonly encoding this protein.

RNA homologous to these genes has been detected in seeds but is not present at detectable levels in seed pod tissue or in vegetative plant tissues. The 12S storage protein mRNA has been detected in seeds that are 7 days old or older. No RNA is detectable in seeds that are younger than 7 days. This message can be detected in every time point sampled after 7 days, which includes seeds up to 15-16 days old, which are still located within the seed pod but which have brown seed coats and have undergone desiccation. Transcripts from both genes studied in this work appear at approximately 7-8 days but differ in the length of the period in which the mRNA can be detected. This difference could be due to cessation of transcription at different times or to differential stabilities of the messages.

The *Arabidopsis* 12S storage protein genes have a structure which is similar to that of other 12S storage protein genes which have been sequenced. They encode a protein which has a sequence homologous with those of the 12S storage proteins of *Brassica napus* and *Pisum sativum* and presumably is processed to mature α and β polypeptides in a similar manner. The locations of the introns in the *Arabidopsis* gene which has been sequenced are in identical positions within the protein sequence as those found in the legumin gene of *P. sativum* (Lycett *et al.*, 1984). All of the introns found in the *Arabidopsis* gene are less than 130 nucleotides in length although each intron is longer than its counterpart in *P. sativum*. No information is available about the intron positions in the *B. napus* gene because the sequence was determined from a cDNA clone. A partial sequence of a genomic clone for the soybean (*Glycine max*) 12S storage protein

gene demonstrates the presence of an intron in this gene in the same position as the third intron of the *Arabidopsis* and *P. sativum* genes (Marco *et al.*, 1984). The intron in the soybean gene is 625 nucleotides in length, making it by far the longest of any of the 12S storage protein gene introns. The *Arabidopsis* gene is more compact than the gene from *Pisum sativum*, but the difference in length is due to a smaller amount of coding sequence which encodes a smaller protein rather than to a reduction in the size or number of introns as is the case with the *Adh* gene of *Arabidopsis* relative to the maize *Adh* genes (Chang and Meyerowitz, 1986).

By comparing the amino acid sequence of the *Brassica napus* and *Pisum sativum* 12S storage proteins Simon *et al.* (1985) identifed two regions which contained large insertions in one protein relative to the other. The sequence of the *Arabidopsis* 12S protein gene reveals that there are small insertions of amino acid sequence present in the *Arabidopsis* protein in these same two locations, thus confirming the hypothesis of Simon *et al.* that these might represent sites in the protein which were tolerant of variation. It is important to note, however, that although the insertions in the *Arabidopsis* protein are substantially different in size from those found in the other proteins, the amino acid compositions are comparable to those of the insertions located in similar locations in the other proteins. This fact may indicate some constraint on these regions which will limit the amount of manipulation of amino acid sequence that they will tolerate. It will be interesting to see if the 12S storage proteins from other plants show sequence variation in these same regions and if so to what extent the variation will include changes in amino acid composition as well as in sequence length.

Acknowledgements

This work was supported by grant number PCM-8408504 from the National Science Foundation to E.M.M.

References

Chang, C. and E.M. Meyerowitz (1986) Molecular cloning and DNA sequence of the *Arabidopsis thaliana* alcohol dehydrogenase gene. *Proc. Natl. Acad. Sci. USA*, in press.

Croy, R.R.D., G.W. Lycett, J.A. Gatehouse, J.N. Yarwood and D. Boulter (1982) Cloning and analysis of cDNAs encoding plant storage protein precursors. *Nature* **295**: 76-78.

Davis, R.W., D. Botstein and J.R. Roth (1980) Advanced bacterial genetics. Cold Spring Harbor Laboratory: Cold Spring Harbor, New York.

Denhardt, D.T. (1966) A membrane-filter technique for the detection of complementary DNA. *Biochem. Biophys. Res. Commun.* **23**:641-646.

Fischer, R.L. and R.B. Goldberg (1982) Structure and flanking regions of soybean seed protein genes. *Cell* **29**: 651-660.

Garfinkel, M.D., R.E. Pruitt and E.M. Meyerowitz (1983) DNA sequences, gene regulation and modular protein evolution in the *Drosophila* 68C glue gene cluster. *J. Mol. Biol.* **168**: 765-789.

Higgins, T.J.V. (1984) Synthesis and regulation of major proteins in seeds. *Ann. Rev. Plant Physiol.* **35:** 191-221.

Hinnebusch, A.G. (1984) Evidence for translational regulation of the activator of general amino acid control in yeast. *Proc. Natl. Acad. Sci. USA* **81:** 6442-6446.

Kahana, C. and D. Nathans (1985) Nucleotide sequence of murine ornithine decarboxylase mRNA. *Proc. Natl. Acad. Sci. USA* **82:** 1673-1677.

Lycett, G.W., R.R.D Croy, A.H. Shirsat and D. Boulter (1984) The complete sequence of a legumin gene from pea (*Pisum sativum* L.). *Nuc. Acids Res.* **12:** 4493-4506.

Lycett, G.W., R.R.D. Croy, A.H. Shirsat, D.M. Richards and D. Boulter (1985) The 5'-flanking regions of three pea legumin genes: comparison of the DNA sequences. *Nuc. Acids Res.* **13:** 6733-6743.

Marco, Y.A., V.H. Thanh, N.E. Tumer, B.J. Scallon and N.C. Nielsen (1984) Cloning and structural analysis of DNA encoding an A_2B_{1a} subunit of glycinin. *J. Biol. Chem.* **259:** 13436-13441.

Maxam, A.M. and W. Gilbert (1980) Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol.* **65:** 499-560.

Meinke, D.W. and I.M. Sussex (1979) Embry-lethal mutants of *Arabidopsis thaliana*. *Dev. Biol.* **72:** 50-61.

Meyerowitz, E.M. and C.H. Martin (1984) Adjacent chromosomal regions can evolve at very different rates: evolution of the *Drosophila* 68C glue gene cluster. *J. Mol. Evol.* **20:** 251-264.

Pernollet, J.-C. (1978) Protein bodies of seeds: ultrastructure, biochemistry, biosynthesis and degradation. *Phytochemistry* **17:** 1473-1480.

Pruitt, R.E. and E.M. Meyerowitz (1986) Characterization of the genome of *Arabidopsis thaliana*. *J. Mol. Biol.*, in press.

Redei, G.P. (1970) *Arabidopsis thaliana* (L.) Heynh. A review of the genetics and biology. *Biblio. Genet.* **20:** 1-151.

Rigby, P.W.J., M. Dieckmann, C. Rhodes and P. Berg (1977) Labeling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase I. *J. Mol. Biol.* **113:** 237-251.

Simon, A.E., K.M. Tenbarge, S.R. Scofield, R.R. Finkelstein and M.L. Crouch (1985) Nucleotide sequence of a cDNA clone of *Brassica napus* 12S storage protein shows homology with legumin from *Pisum sativum*. *Plant Mol. Biol.* **5:**191-201.

Thireos, G., M.D. Penn and H. Greer (1984) 5' untranslated sequences are required for the translational control of a yeast regulatory gene. *Proc. Natl. Acad. Sci. USA* **81:** 5096-5100

Figure 1. Restriction maps of the regions containing the 12S storage protein genes from the two *Arabidopsis* strains Landsberg and Columbia. All sites are shown for the restriction enzymes Eco RI, HindIII, Sal I and XhoI. The arrows beneath the maps indicate the locations of the storage protein genes and the direction of transcription. The locations of the genes was determined by hybridization of a cDNA clone from *Brassica napus* to the cloned DNA and by DNA sequencing. The hatched boxes represent the regions sequenced. The locations of the fragments contained in the subclones nAt1512, sAt2101 and sAt2105 are also indicated.

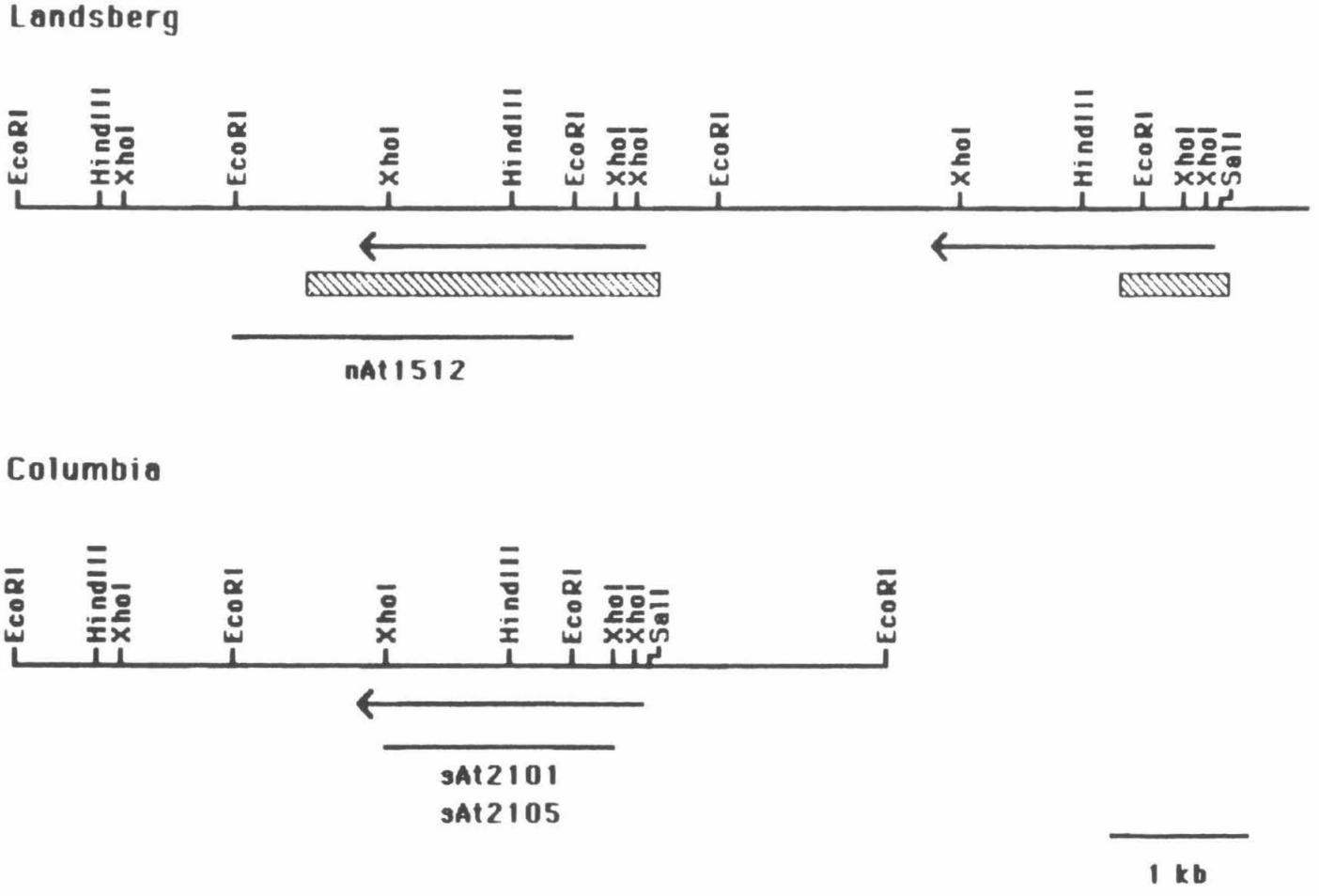

Landsberg
EcoRI
HindIII
XhoI
EcoRI
XhoI
HindIII
EcoRI
XhoI
XhoI
EcoRI
XhoI
HindIII
EcoRI
XhoI
XhoI
SalI
nAt1512
Columbia
EcoRI
HindIII
XhoI
EcoRI
XhoI
HindIII
EcoRI
XhoI
XhoI
SalI
EcoRI
sAt2101
sAt2105
1 kb

Figure 2. Determination of transcription direction of both group I and group II genes. RNA prepared from four seed pods was denatured and fractionated on a formaldehyde-agarose gel and transferred to a nitrocellulose filter. This filter was separated into four strips each of which was allowed to hybridize with a single-stranded RNA probe which was prepared from the clone indicated at the top of the figure. Sp6 and T7 denote probes made with Sp6 or T7 RNA polymerases, respectively.


sAt2101
sAt2105
At1519/Sp6
At1519/T7

Figure 3. Restriction map of the region containing the group II genes. All known restriction sites for the enzymes Bgl II, EcoRI and SacI are shown. In addition, the HindIII sites which bound the plasmid subclone nAt1510 are shown. The arrows represent the locations and directions of transcription of the regions which encode the 650 nucleotide RNA.

EcoRI
BglII
BglII
SacI
BglII
EcoRI
HindIII
SacI
HindIII
EcoRI
EcoRI
nAt1511
nAt1510
1 kb

Figure 4. Detailed restriction maps of the plasmid subclones nAt1510 and nAt1511. All restriction sites for the enzymes BglII, Eco RI, Hinc II, HindIII, PstI, SacI, SpeI and SphI are presented. The regions which hybridize with labeled cDNA from seeds are again denoted by arrows which also show the directions of transcription.

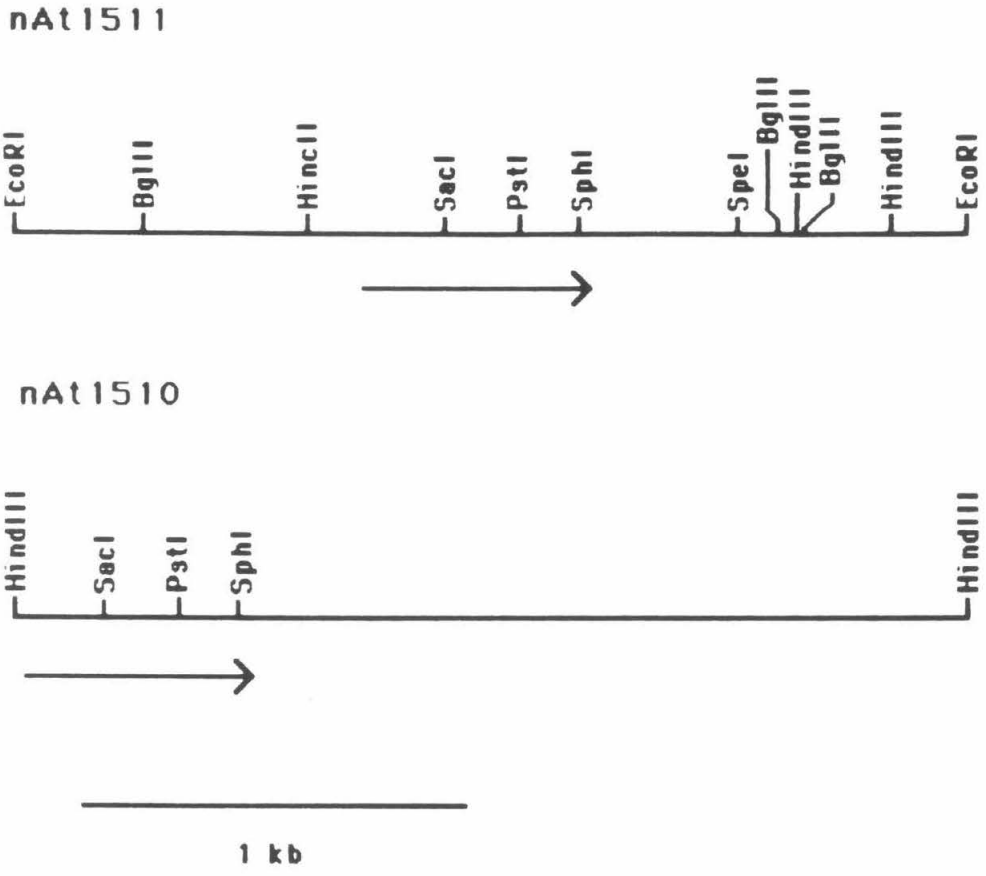
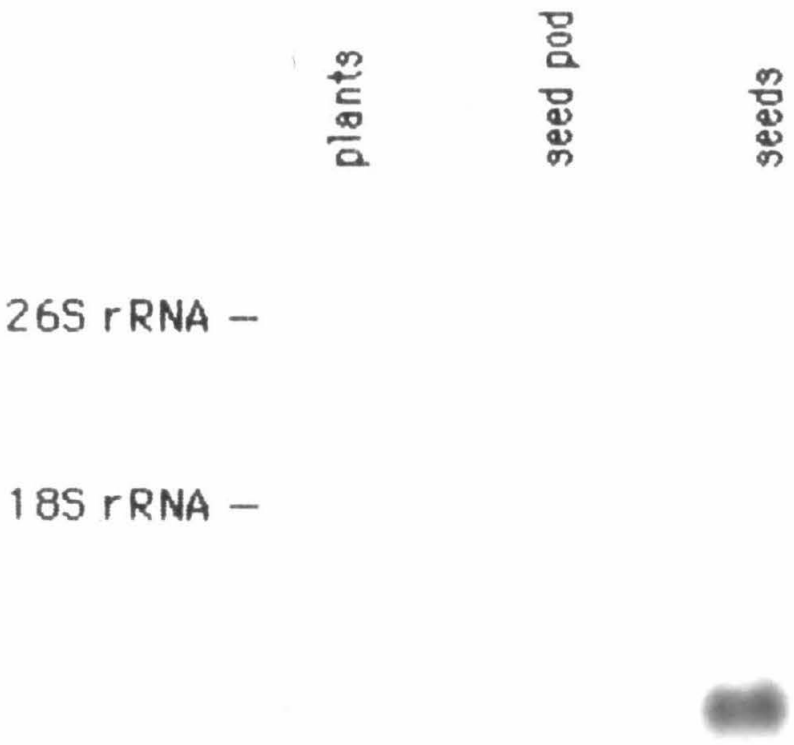
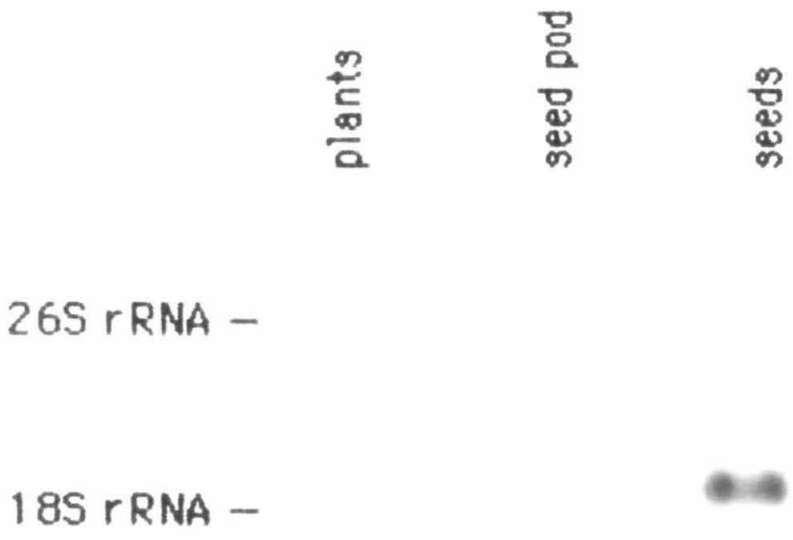

nAt1511
EcoRI
BglII
HincII
SacI
PstI
SphI
SpeI
BglII
HindIII
BglII
HindIII
EcoRI
nAt1510
HindIII
SacI
PstI
SphI
HindIII
1 kb

Figure 5. Tissue specific expression of both group I and group II genes. RNA was prepared from 50 whole plants, 2 seed pods with the seeds removed (7-8 days old) or 10 seeds (also 7-8 days old) denatured, the samples divided equally into two halves and fractionated on the two sides of a formaldehyde-agarose gel. Following electrophoresis the RNA was blotted to duplicate nitrocellulose filters and one filter allowed to hybridize with ^{32}P-labeled DNA from nAt1512 (Figure 5a) and the other with ^{32}P-labeled DNA from nAt1510 (Figure 5b). After autoradiography the probe was removed from one of these filters and the filter reprobed with λbAt106 which contains an rDNA repeat from *Arabidopsis* (Figure 5c).


plants
seed pod
seeds
26S rRNA —
18S rRNA —


plants
seed pod
seeds
26S rRNA –
18S rRNA –

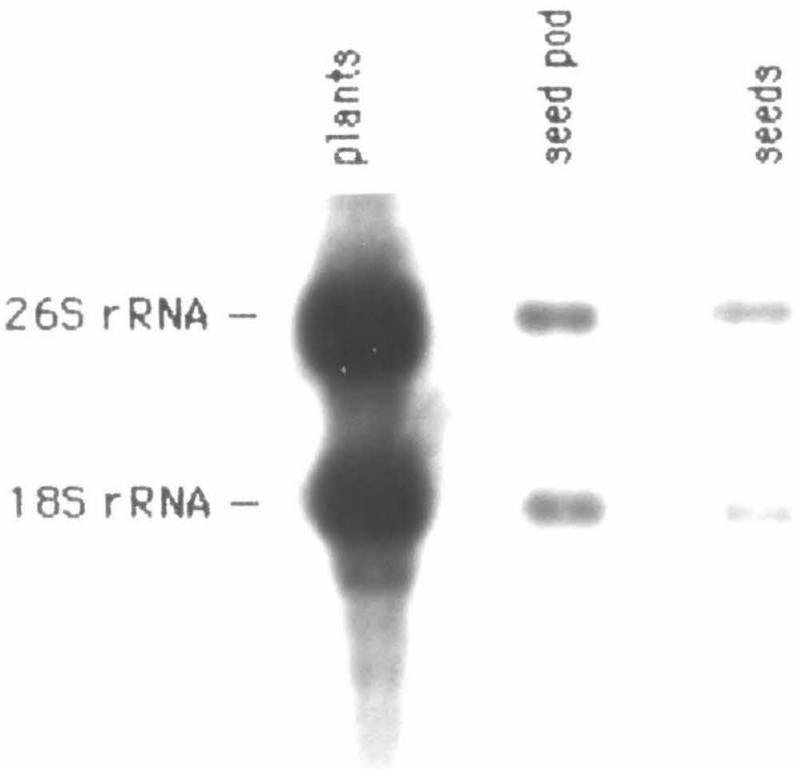

plants
seed pod
seeds
26S rRNA –
18S rRNA –

Figure 6. Determination of the times in seed development when each of the groups of genes is expressed. RNA was prepared from 2 seed pods of each of the ages shown at the top of the figure, denatured and separated on a formaldehyde-agarose gel. The RNA from the unfertilized buds was prepared from approximately 25 unfertilized flowers. The RNA was blotted to a nitrocellulose filter and the filter probed successively with nAt1510 (Figure 6a), nAt1512 (Figure 6b) and λbAt106 (Figure 6c).

Unfertilized
1 - 2
3 - 4
5 - 6
7 - 8
9 - 10
11 - 12
13 - 14
15 - 16
days after fertilization
26S rRNA
18S rRNA

Unfertilized
1-2
3-4
5-6
7-8
9-10
11-12
13-14
15-16
days after fertilization
26S rRNA
18S rRNA

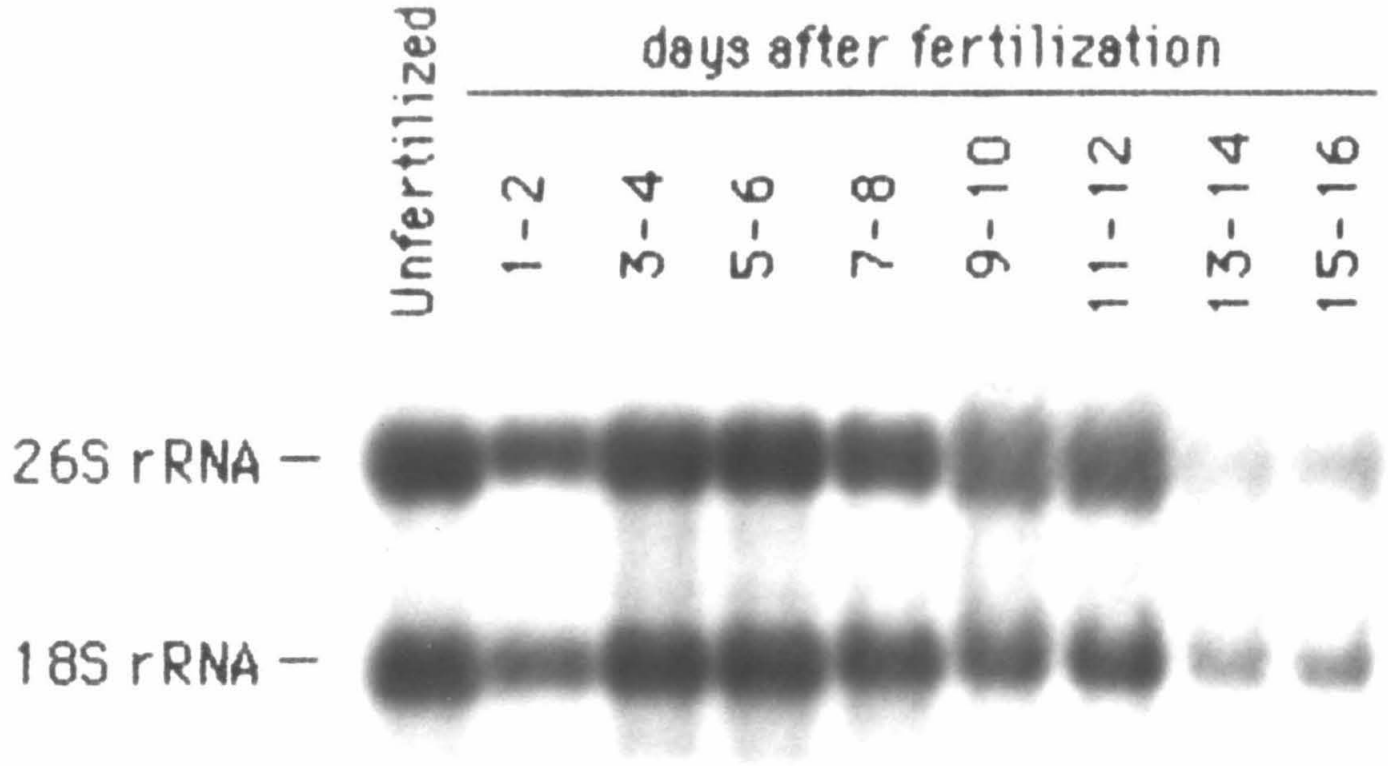

Unfertilized
days after fertilization
1-2
3-4
5-6
7-8
9-10
11-12
13-14
15-16
26S rRNA
18S rRNA

Figure 7. The nucleotide sequence of a group II gene. The sequence is from the plasmid nAt1511 and is presented in the direction the gene is transcribed. The protein sequence of the 148 amino acid open reading frame is also presented below the sequence.

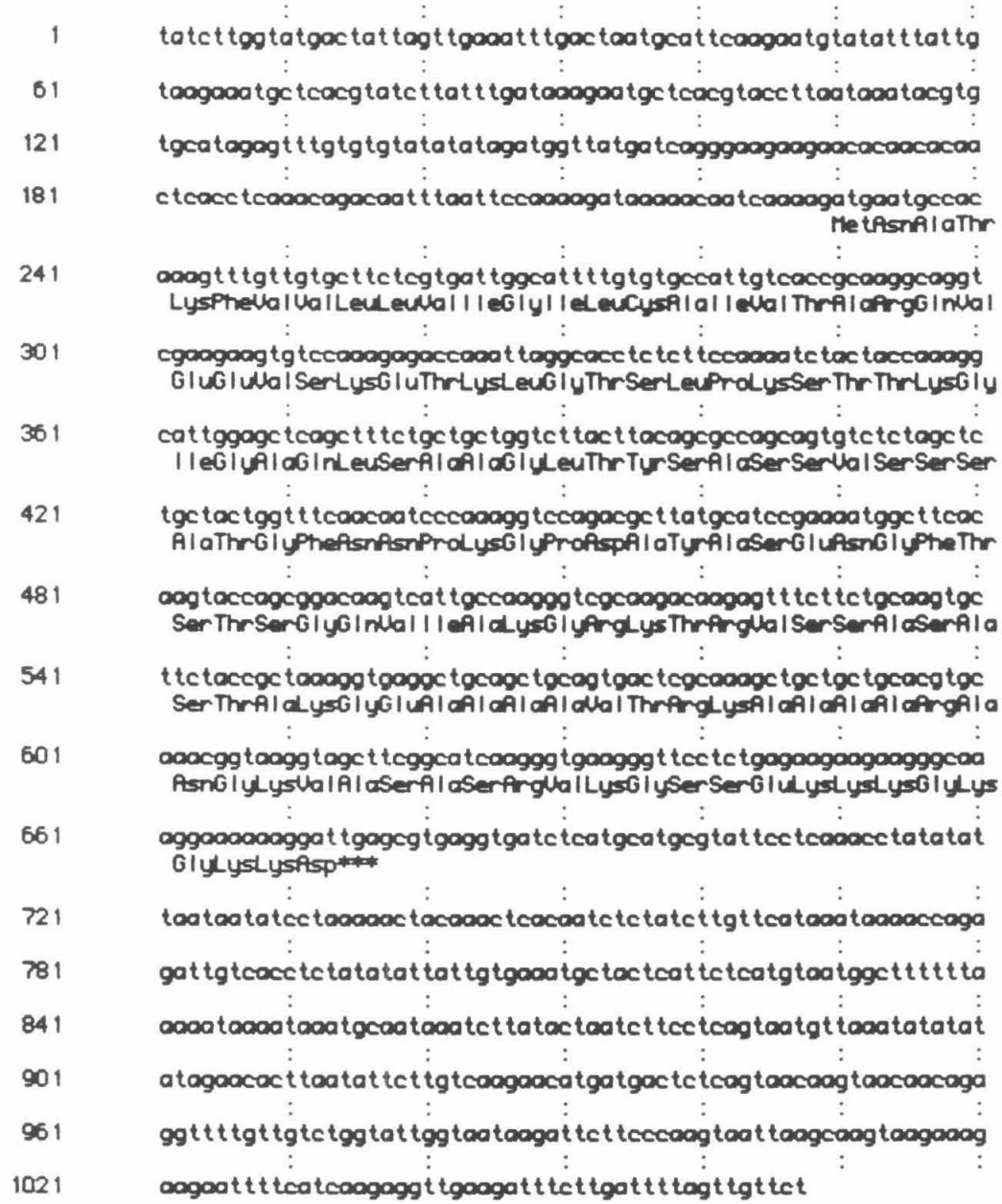
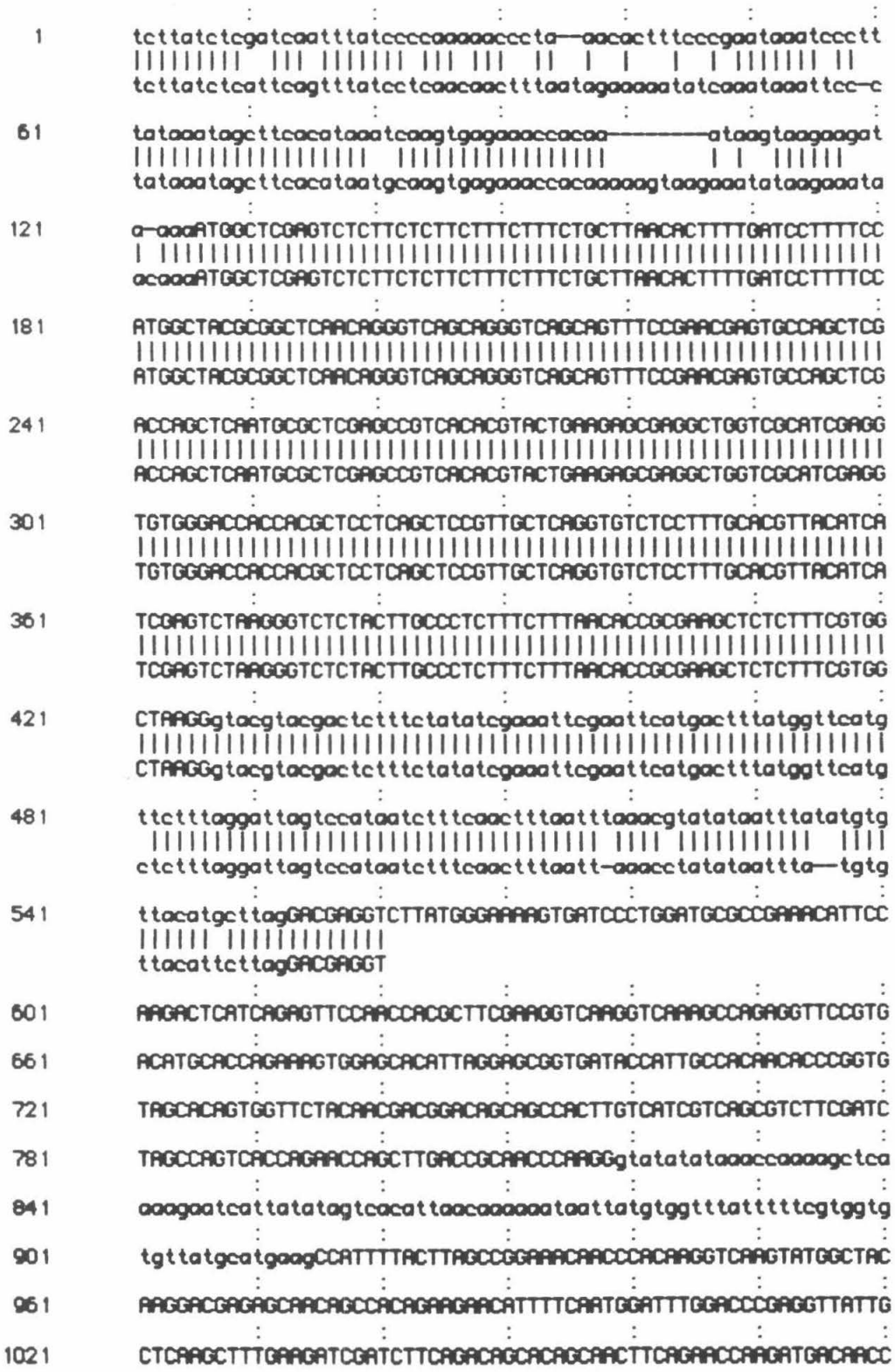
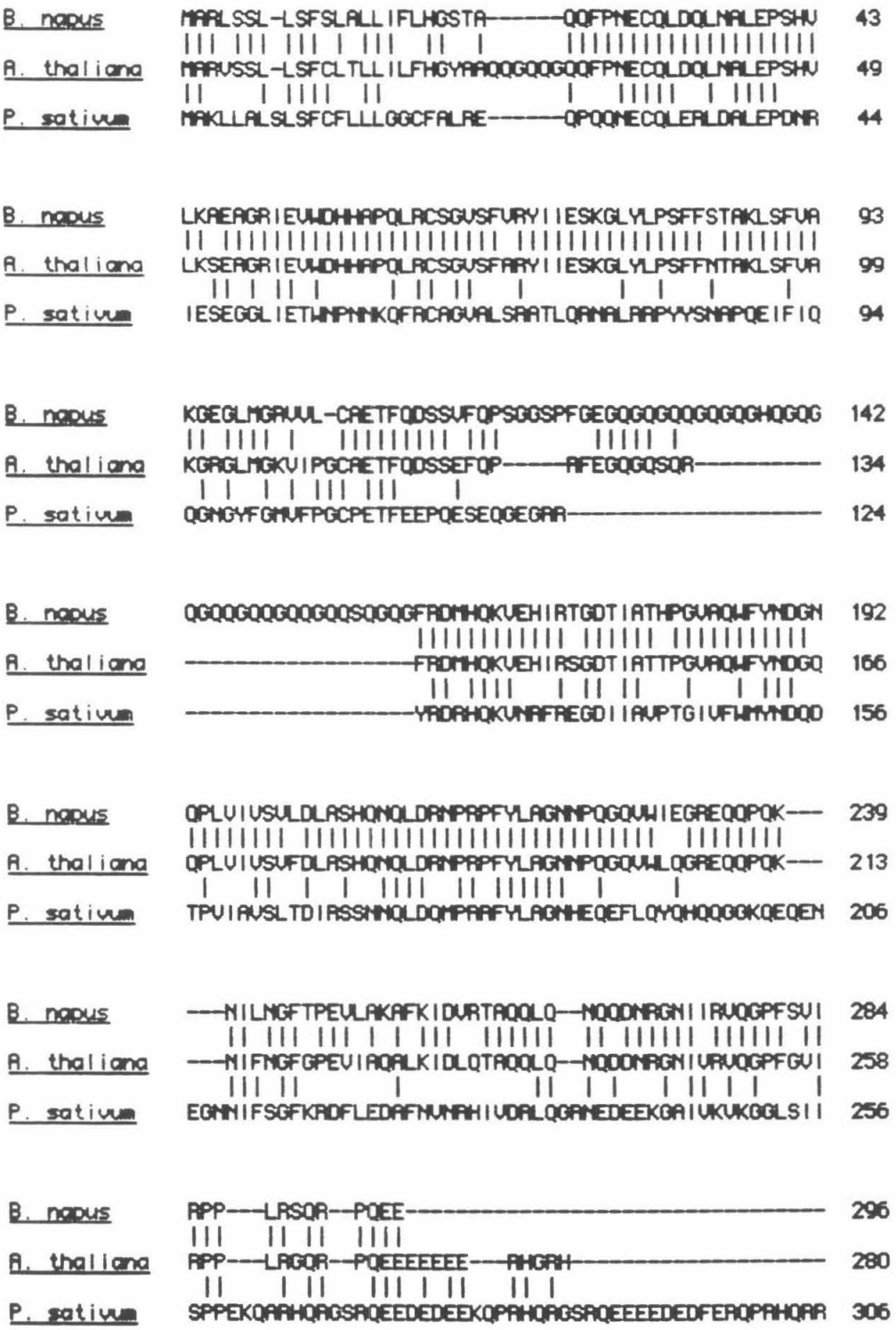

1 tatcttggtatgactattagttgaaatttgactaatgcattcaagaatgtatatttattg
61 taagaaatgctcacgtatcttatttgataaagaatgctcacgtaccttaataaatacgtg
121 tgcatagagtttgtgtgtatatatagatggttatgatcagggaagaagaacacaacacaa
181 ctcacctcaaacagacaatttaattccaaaaagataaaaacaatcaaaagatgaatgccac
MetAsnAlaThr
241 aaagtttgttgtgcttctcgtgattggcattttgtgtgccattgtcaccgcaaggcaggt
LysPheValValLeuLeuValIleGlyIleLeuCysAlaIleValThrAlaArgGlnVal
301 cgaagaagtgtccaaagagaccaaattaggcacctctcttccaaaatctactaccaaagg
GluGluValSerLysGluThrLysLeuGlyThrSerLeuProLysSerThrThrLysGly
361 cattggagctcagctttctgctgctggtcttacttacagcgccagcagtgtctctagctc
IleGlyAlaGlnLeuSerAlaAlaGlyLeuThrTyrSerAlaSerSerValSerSerSer
421 tgctactggtttcaacaatcccaaaggtccagacgcttatgcatccgaaaatggcttcac
AlaThrGlyPheAsnAsnProLysGlyProAspAlaTyrAlaSerGluAsnGlyPheThr
481 aagtaccagcggacaagtcattgccaagggtcgcaagacaagagtttcttctgcaagtgc
SerThrSerGlyGlnValIleAlaLysGlyArgLysThrArgValSerSerAlaSerAla
541 ttctaccgctaaaggtgaggctgcagctgcagtgactcgcaaagctgctgctgcacgtgc
SerThrAlaLysGlyGluAlaAlaAlaAlaValThrArgLysAlaAlaAlaAlaArgAla
601 aaacggtaaggtagcttcggcatcaagggtgaagggttcctctgagaagaagaagggcaa
AsnGlyLysValAlaSerAlaSerArgValLysGlySerSerGluLysLysLysGlyLys
661 aggaaaaaaggattgagcgtgaggtgatctcatgcatgcgtattcctcaaacctatatat
GlyLysLysAsp***
721 taataatatcctaaaaactacaaactcacaatctctatcttgttcataaataaaaccaga
781 gattgtcacctctatatattattgtgaaatgctactcattctcatgtaatggctttttta
841 aaaataaaataaatgcaataaatcttatactaatcttcctcagtaatgttaaatatatat
901 atagaacacttaatattcttgtcaagaacatgatgactctcagtaacaagtaacaacaga
961 ggttttgttgtctggtattggtaataagattcttcccaagtaattaagcaagtaagaaag
1021 aagaattttcatcaagaggttgaagatttcttgattttagttgttct

Figure 8. The nucleotide sequence of the *Arabidopsis* 12S storage protein gene. The sequence of the left gene of Figure 1 was determined in its entirety. Protein coding sequences are denoted by upper case letters and non-coding portions by lower case letters. The sequence of the 5' end and flanking region of the right gene is also presented below the sequence of the complete gene. A vertical bar (|) indicates homology between the two sequences at a given position. A dash (-) indicates a gap introduced to maximize sequence alignment.




1081 GTGGAAACATTGTCCGAGTCCAAGGACCGTTCGGTGTCATTAGGCCGCCTTTGAGGGGCC
1141 AGAGACCTCAGGAGGAGGAAGAAGAAGAAGGACGACATGGACGACACGGTAATGGCTTAG
1201 AGGAGACCATCTGCAGCGCCAGGTGCACCGATAACCTCGATGACCCGTCTCGTGCTGACG
1261 TGTACAAGCCACAGCTCGGTTACATCAGCACTCTCAACAGTTACGATCTCCCCATTCTTC
1321 GCTTCATCCGTCTCTCAGCCCTCCGTGGATCTATCCGTCAAgtaagtaaacataaatatt
1381 atgttactataacctagtaaaatatgcatgcctgatgcatgttaatatgtccatttctat
1441 atttaaaacatgactcttgaaaacgtgtgtgggtgtagAACGCAATGGTGCTTCCACAGTGG
1501 AACGCAAACGCGAACGCTATTCTTTACGAGACAGACGGGGAAGCCCAAATCCAGATCGTA
1561 AACGACAATGGTAACAGAGTGTTTGACGGACAAGTCTCTCAAGGACAGCTCATAGCCGTA
1621 CCACAAGGTTTCTCGGTGGTGAAACGCGCAACAAGCAACCGATTCCAGTGGGTTGAGTTC
1681 AAAACAAACGCTAACGCGCAAATCAACACTCTGGCGGGACGAACCTCAGTCTTGAGAGGT
1741 TTACCACTTGAAGTCATAACCAATGGGTTCCAAATCTCACCCGAAGAAGCAAGGAGGGTC
1801 AAGTTCAACACGCTCGAGACCACTTTGACTCACAGCAGTGGCCCAGCTAGCTACGGAAGG
1861 CCAAGAGTGGCTGCAGCTtaagagcttaaaaccgcggcttaacaatgaaccgctactgta
1921 aaaggaagttaaatagtacgtagtaataataataatgtacgaaaatgtgactagttttgt
1981 tgaggtttacctgtaaaatgcaactcctttttctgaataaaatcttttcaattttcgatca
2041 agttaatacaaatctaggtctaaattaggttcttaatcatagagactagttctgatttttt
2101 atgatttaatacatttgaatcatcatattatttttatataataatccaatattaacattag
2161 acaagtcgccaaaatattgtcatgcttaacaaatttatattacctcatttttctttatcta
2221 tttataatacatcaaatgctttaatttttaatttcaaatatctaatttaatccgtgc

Figure 9. Protein sequences of three 12S storage protein genes. The sequences of *Brassica napus* and *Pisum sativum* are from the papers by Simon *et al.* (1985) and Lycett *et al.* (1984). The numbers on the right reflect amino acid position in each particular sequence. A vertical bar (|) denotes a homologous position. A dash (-) represents a gap introduced into the protein sequence to maximize alignment.


B. napus
A. thaliana
P. sativum

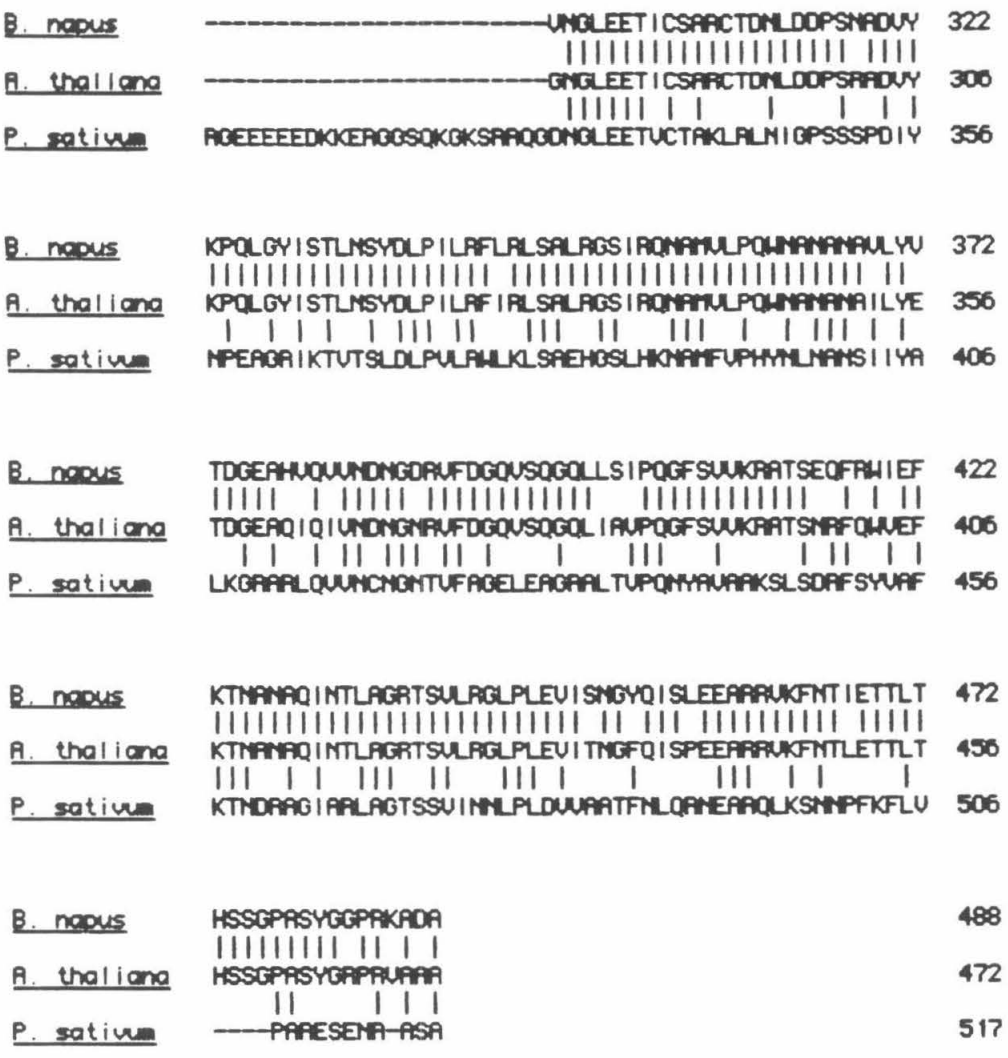
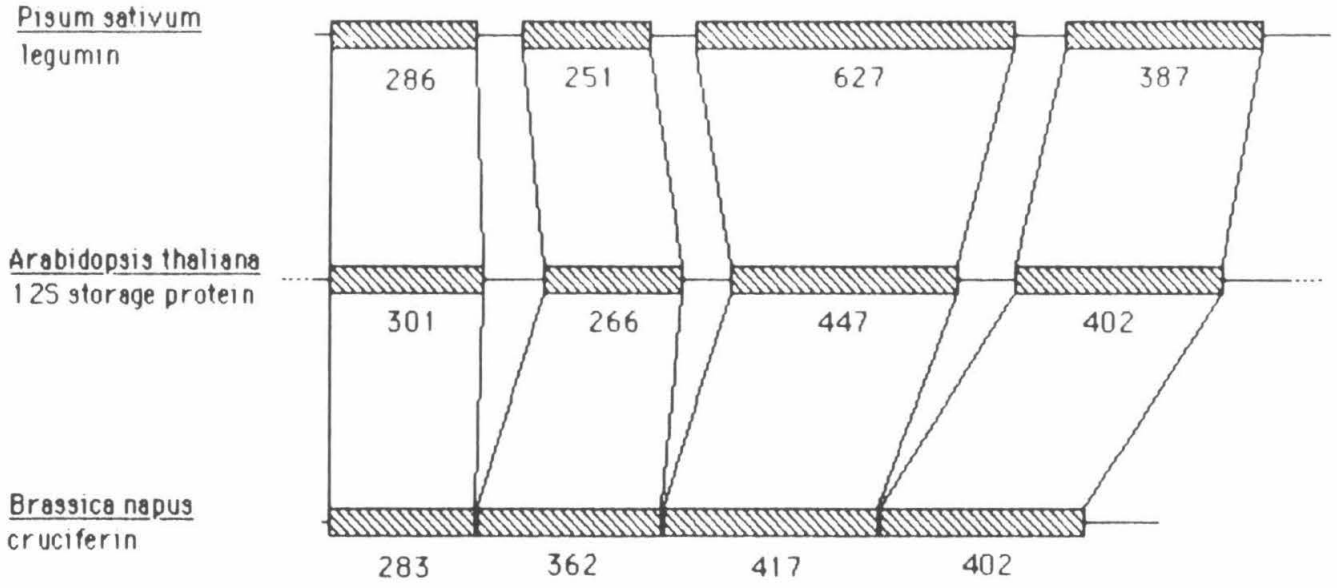

B. napus
A. thaliana
P. sativum
322
306
356
372
356
406
422
406
456
472
456
506
488
472
517

Figure 10. Diagrammatic representation of the structures of the 12S storage protein genes encoding the proteins presented in Figure 9. The boxes are exons and the thin lines non-coding regions (introns and 5' or 3' untranslated regions). The numbers are the number of nucleotides in each exon. The *Brassica napus* gene sequence was determined from a cDNA clone and, as such, no information is available concerning intron positions or sizes. The lines connecting the genes connect homologous positions in the *Arabidopsis* and *Brassica napus* genes allowing comparison of the changing sizes of the different regions of the protein.

12S storage protein gene structure



Appendix A

DNA sequences, Gene Regulation and Modular Protein Evolution in the *Drosophila* 68C Glue Gene Cluster

(reprinted from the Journal of Molecular Biology **168:** 765-789.)

J. Mol. Biol. (1983) **168**, 765–789

DNA Sequences, Gene Regulation and Modular Protein Evolution in the *Drosophila* 68C Glue Gene Cluster

MARK D. GARFINKEL, ROBERT E. PRUITT AND ELLIOT M. MEYEROWITZ†

Division of Biology, California Institute of Technology
Pasadena, Calif. 91125, U.S.A.

(Received 16 February 1983, and in revised form 31 March 1983)

The 68C locus of the *Drosophila melanogaster* polytene chromosomes contains the structural genes for three glue polypeptides (sgs-3, sgs-7 and sgs-8) synthesized in the larval salivary glands during the third larval instar. When the messenger RNAs for the glue polypeptides are being synthesized, the locus is puffed; the puff regresses in response to the steroid hormone ecdysterone. The three 68C glue mRNAs are coded in a gene cluster of less than 5000 base-pairs, and are expressed co-ordinately. In the experiments described here we show that the co-ordinate expression of these RNAs does not result from amplification of the puff DNA, nor is it associated with puff DNA rearrangement. We also report the nucleotide sequence of 6751 base-pairs of genomic DNA that includes the entire gene cluster, and describe coding and non-coding sequences with possible regulatory roles. In addition, we deduce the amino acid sequences of the primary translation products of the glue mRNAs, and show that the glue proteins form a diverged gene family. The members of the family all contain an amino-terminal hydrophobic block of amino acids, which is absent in the mature, secreted glue proteins, and a cysteine-rich carboxy-terminal module. sgs-3 differs from sgs-7 and sgs-8 by containing a third module between the other two, comprised largely of tandem repeats of the five amino acids Pro-Thr-Thr-Thr-Lys.

1. Introduction

During the third larval instar, the major function of the *Drosophila melanogaster* salivary glands is the production of a set of secreted polypeptides (Zhimulev & Kolesnikov, 1975). There are at least eight of these (Crowley *et al.*, 1983), synthesized in the cytoplasm of the salivary gland cells throughout the instar and secreted into the lumen of the gland near the end of the larval stage. At the time of puparium formation the luminal contents are expelled, and they set to form a glue that fixes the puparium to its substrate for the duration of the pupal period (Lane *et al.*, 1972). A group of about ten puffs, or sites of highly active transcription, are present on the giant polytene chromosomes of the salivary gland cells when glue proteins are being synthesized; they disappear toward the end of the third larval instar, when glue synthesis terminates. These are known as the intermolt puffs (Ashburner, 1972). Genetic, cytogenetic and molecular

† Author to whom all correspondence should be sent.

0022–2836/83/240765–25 $03.00/0

mapping experiments have shown that at least four of these puffs contain structural genes for at least six of the glue polypeptides (Korge, 1975, 1977; Akam *et al.*, 1978; Velissariou & Ashburner, 1980, 1981; Crowley *et al.*, 1983). Other experiments have shown that the regression of the intermolt puffs is a consequence of an increase in the titer of the steroid hormone ecdysterone in the larval hemolymph several hours before pupariation (Ashburner, 1973). The regression of one of these puffs, that at 68C on the left arm of the third chromosome, appears to result directly from binding of the hormone (presumably through the mediation of a steroid receptor protein) to the puff (Gronemeyer & Pongs, 1980).

The molecular cloning of the 68C puff genomic DNA and of DNA complementary to the puff-encoded RNAs has been accomplished. Analysis using the cloned DNA has shown that one 5000 base-pair region of the puff DNA codes for three different polyadenylated messenger RNAs, all found only in third larval instar salivary glands, and appearing and disappearing co-ordinately (Meyerowitz & Hogness, 1982). Each of the 68C RNAs, designated the group II, group III and group IV RNAs, is translated to a different salivary gland glue polypeptide, sgs-8, sgs-7 and sgs-3, respectively (Crowley *et al.*, 1983).

There are at least two features of the regulation of the 68C puff gene cluster that must be understood: the co-ordinate control of the three different RNAs, and the action of ecdysterone in puff regression. In the experiments described here, we test several hypotheses for the mechanism of co-ordinate regulation of the 68C glue RNAs, and in so doing find that the 68C glue proteins are evolutionarily related to each other in an unusual way. In addition, DNA sequence information obtained in these experiments constrains the possible types of regulatory DNA sequences that can be considered as important in co-ordinate regulation of the puff RNAs.

2. Materials and Methods

(a) *Insect culture*

Adult flies of the *D. melanogaster* third chromosome homozygous strain OR16f (Meyerowitz & Hogness, 1982) were reared in milk bottles or in population cages similar to the design of Elgin & Miller (1978) at 22 °C. They were fed standard cornmeal–agar food that was supplemented with live yeast paste. Eggs were laid on food coated with yeast in plastic trays. The trays were covered with tight-fitting plastic boxes. Larvae were refed live yeast and were watered daily. Late third instar larvae were collected on days 5 and 6 after egg deposition.

(b) *Isolation of third instar salivary gland nucleic acids*

Third instar larvae were washed from the trays and boxes with cold distilled water. Food particles were removed by floating the larvae in 20% (w/v) sucrose. The clean larvae were then washed in Robb's (1969) PBS (phosphate-buffered saline), and crushed between metal rollers. Salivary glands and carcasses were collected on a fine-mesh Nitex (Tetko, Inc.) screen. Glands were separated from carcasses by filtration through a coarse-mesh Nitex screen. Salivary glands were collected in a plastic beaker. Fat bodies were removed from the glands by repeatedly allowing them to sediment at unit gravity. Gut, Malpighian tubules, and other tissues were removed from the glands by centrifugation

through 32% Ficoll (Sigma) in Robb's PBS. Ficoll was removed by washing the glands with Robb's PBS. They were judged to be greater than 70% pure salivary glands.

RNA was extracted from the glands by dissolving the tissue in 0·1 M-Tris·HCl (pH 8·0), 0·2 M-NaCl, 0·1 M-EDTA, 0·5% (w/v) sodium dodecyl sulfate. Repeated phenol, phenol/chloroform, and ether extractions were performed. The nucleic acids were precipitated with ethanol, washed and precipitated with ethanol again. Several mg of RNA were recovered from several hundred mg of tissue. RNA was stored in 10 mM-sodium acetate (pH 5·0) at −80 C.

Poly(A)$^+$ RNA was obtained from salivary gland RNA by oligo(dT)-cellulose chromatography as described by Meyerowitz & Hogness (1982).

DNA was prepared from the glands by a modification of the procedure used by Meyerowitz & Hogness (1982) to obtain adult fly DNA. About 100 mg salivary gland tissue in Robb's PBS was treated with 1 ml 15% sucrose, 50 mM-Tris·HCl (pH 8·0), 50 mM-EDTA. The tissue was spun briefly in a hand-driven centrifuge and the supernatant discarded. One ml of 0·12 M-sucrose, 150 mM-Tris·HCl (pH 8·5), 75 mM-EDTA, 0·75% sodium dodecyl sulfate was added to the tissue. Five μl 25% (v/v) diethyl pyrocarbonate in ethanol were added and the glands were lysed in a 2 ml Ten-Broeck homogenizer. The mixture was transferred to a 1·5 ml capped plastic tube, 65 μl 8 M-potassium acetate were added, and the mixture allowed to stand in ice for 15 min. Precipitated debris and potassium dodecyl sulfate were removed by a 10-min spin in a micro-centrifuge. The supernatant was transferred to two 1·5-ml tubes and 1·1 ml ethanol were added to each tube. Nucleic acids were pelleted in the hand-driven centrifuge, rinsed twice with 70% (v/v) ethanol and air-dried. Each pellet was resuspended in 20 μl 10 mM-Tris·HCl (pH 8·0), 1 mM-EDTA, and 100 ng pancreatic RNase A (a gift from D. Ridge) were added.

(c) *General DNA and recombinant DNA techniques*

Plasmids were grown in *Escherichia coli* HB101 using M9–Casamino acids supplemented with uridine as medium (Norgard *et al.*, 1979). Chloramphenicol amplification was sometimes used. Plasmid purification by CsCl/ethidium bromide gradient centrifugation was performed as described (Meyerowitz & Hogness, 1982).

Recombinant λ phage were propagated on *E. coli* K802 and were purified by standard methods (Maniatis *et al.*, 1978; Meyerowitz & Hogness, 1982). Phage DNA was extracted by the rapid formamide method "A" of Davis *et al.* (1980).

National Institutes of Health guidelines were followed for the P1-EK1 level containment of recombinant DNA-bearing organisms.

Preparation of *Drosophila* genome blot filters, nick-translation, and filter hybridization were performed as described by Meyerowitz & Hogness (1982).

(d) *DNA sequence determination by partial chemical cleavage*

(i) *End-labeling DNA*

After restriction enzyme digestion of plasmid DNA, one of three methods was used: for 5′ protruding restriction site termini the 5′ ends were labeled by dephosphorylation with calf intestinal alkaline phosphatase and subsequent rephosphorylation with [γ-^{32}P]ATP and T4 polynucleotide kinase (Maxam & Gilbert, 1980); 3′ ends were labeled by incubating the DNA in 20 μM each of 3 unlabeled deoxynucleoside triphosphates, one [α-^{32}P]deoxynucleoside triphosphate, and *E. coli* DNA polymerase I Klenow fragment. For 3′ protruding restriction sites the 3′ ends were labeled with [α-^{32}P]CTP and terminal deoxynucleotidyl transferase as described by Roychoudhury & Wu (1980).

Fragments labeled at one end were obtained by digestion with a second restriction enzyme.

(ii) *Gel purification of labeled DNA*

The ^{32}P-labeled DNAs were resolved on polyacrylamide gels crosslinked with *N,N′*-bis-acrylyl-cystamine (Bio-Rad). Fragments were located by autoradiography, excised from

the gel, and were released from the gel matrix by adding 2-mercaptoethanol to a concentration of 50% (v/v). The disulfide crosslinks are reduced after 0·5 to 2 h at room temperature. Nine ml of 0·1 M-Tris·HCl (pH 7·5), 0·1 M-NaCl were added and each gel slice was homogenized by thorough vortex mixing. The viscous mixes were incubated with 0·2 ml Whatman DE52 DEAE-cellulose resin, which was kept suspended by constant agitation. After several hours at room temperature, the radioactive resins were pelleted in a table top centrifuge, washed twice with 10 ml of 0·1 M-Tris·HCl (pH 7·5), 0·1 M-NaCl to remove acrylamide residue, and were packed into small columns. DNA was eluted from the resin with three 1 ml volumes of 0·1 M-Tris·HCl (pH 7·5), 1 M-NaCl. Fine resin particles were pelleted by centrifugation, and the DNA precipitated from the supernatant by adding 10 μg yeast transfer RNA and 9 ml ethanol, followed by incubation overnight at −20°C. Recoveries generally exceeded 90%.

(iii) *Limited modification of bases*

The Maxam & Gilbert (1977,1980) procedure modified by Smith & Calvo (1980) forms the basis of our sequence determination protocol. Base modification conditions were chosen to enable us to read up to 650 nucleotides from a labeled end. The G + A reaction used only 2 μl 1 M-pyridinium formate in a 22-μl volume at 37°C for 5 min. Hydrazine reactions (C, C + T) were performed in an ice water bath for 15 min. The dimethyl sulfate G reaction was done for 12 to 15 min in an ice water bath using 0·125% (v/v) dimethyl sulfate. Stop solution, ethanol precipitation, and ethanol rinse steps were done as described (Maxam & Gilbert, 1980).

(iv) *Gel electrophoresis and autoradiography*

Sequence gels were 0·36 mm thick, and contained 100 mM-Tris/borate/EDTA (Maxam & Gilbert, 1980) and 50% (w/v) urea. To read nucleotides 1 to 50, a 40-cm long 20% polyacrylamide gel was run at 40 W constant power until the xylene cyanol marker had migrated 12 cm. To read nucleotides 35 to 650, multiple staggered runs on 80-cm long 5% polyacrylamide gels were performed. The gels were run at 2400 to 2800 V constant potential. Repeated loadings on one gel, or several gels run for different times, were used such that the xylene cyanol marker was allowed to migrate 30 cm, 90 cm, 150 cm or 210 cm. This pattern of electrophoresis allowed for alignment of overlapping contiguous sequence.

Gels were transferred from the glass plates to sheets of Whatman 3MM paper or scrap X-ray film, covered with plastic wrap, and autoradiographed. Kodak XR-5 or XAR-5 film was used. Duplicate exposures of the 150 cm run and 210 cm run gels were done with or without DuPont Cronex Lighting-Plus intensifier screens. All autoradiographs were read independently by 2 persons. Discrepancies were resolved by reference to the original films, and by additional sequence determinations. Except for the leftmost 70 nucleotides, which were determined once, every position was assigned on the basis of at least 2 independent experiments. See Fig. 3 for additional details.

(e) *Nuclease mapping of mRNAs*

^{32}P-labeled restriction fragments (about 50,000 cts/min) were mixed with 20 μg yeast tRNA for mock hybridizations, or with 20 μg yeast tRNA and 2 μg salivary gland poly(A)$^+$ RNA. The nucleic acids were precipitated with ethanol, rinsed, and dried. Hybridizations were carried out by dissolving the nucleic acids in 100 μl of 70% deionized formamide, 10 mM-PIPES (pH 6·4), 0·4 M-NaCl, 0·1 mM-EDTA (Casey & Davidson, 1977), heating to 70°C for 10 min and annealing at 50·0°C for several hours. While leaving the hybridizations at 50·0°C, 15-μl portions were removed and diluted into 200-μl portions of nuclease assay buffer. The assay buffer tubes were pre-chilled in ice water baths, and enzyme was already added to the appropriate tubes. Rapid transfer of the hybridization

portion, forceful pipetting, rapid vortex mixing, and immediate transfer to the digestion temperature all ensured that strand displacement was minimized.

The nuclease S_1 reaction was 0·3 M-sodium acetate (pH 4·5), 0·4 M-NaCl, 0·1 mM-zinc acetate, 30 μg heat-denatured salmon sperm DNA/ml at 37°C for 15 min. The reaction was terminated by adding 600 μl ethanol and freezing on solid CO_2. The precipitated nucleic acids were pelleted, washed with 70% ethanol, dried and resuspended in Maxam & Gilbert (1980) sequence gel sample buffer. *Aspergillus oryzae* nuclease S_1 was obtained from Sigma Chemical Co.

Exonuclease VII digestions were carried out in 10 mM-Tris·HCl (pH 7·5), 20 mM-KCl, 10 mM-EDTA (Berk & Sharp, 1978) at 45°C for 45 min. The reaction was terminated by adding 20 μl 3 M-sodium acetate (pH 5·0) and 600 μl ethanol. Nucleic acids were precipitated, washed, and resuspended as described for the nuclease S_1 samples. *E. coli* exonuclease VII was obtained from Bethesda Research Laboratories.

Portions of the initial ^{32}P-labeled restriction fragment were subjected to the partial chemical degradation sequence reactions. The nuclease digests were run alongside the sequence size standards on 5% polyacrylamide/urea gels which were 80 cm long.

(f) *Primer extension sequence determination*

These experiments were done according to the Ghosh *et al.* (1980) method as modified by Snyder *et al.* (1982). ^{32}P-labeled restriction fragment was mixed with 0·5 to 0·8 mg salivary gland RNA, precipitated with ethanol, rinsed, dried and hybridized in 200 μl of 70% formamide as described for nuclease mapping. The hybridization mix was then diluted by adding 1·5 ml oligo(dT) binding buffer. [^{32}P]DNA–poly(A)$^+$ RNA hybrids were recovered using oligo(dT)-cellulose chromatography. The hybrids were eluted, precipitated with ethanol, rinsed, dried and resuspended in reverse transcriptase buffer. Avian myeloblastosis virus reverse transcriptase (a gift from J. Beard) was added and the reaction incubated at 37°C for 3 h. NaOH was added to 0·1 M, and RNA hydrolyzed for 1 h at 37°C. The reaction was neutralized, extracted with phenol, extracted with chloroform, and precipitated twice with ethanol. The complementary DNA was rinsed with ethanol, dried and resuspended in water. The complementary DNA was divided into 5 batches, 4 for the sequence determination reactions and 1 as a standard.

(g) Bal*31 deletion construction*

Two μg of aDm2023 plasmid DNA were linearized by complete digestion with *Xho*I in a 20-μl reaction. After digestion, 12 μl water and 8 μl 5 × *Bal*31 buffer (1 × = 20 mM-Tris·HCl, pH 8·1, 12 mM-$CaCl_2$, 12 mM-$MgCl_2$, 0·6 M-NaCl, 1 mM-EDTA) were added. 0·41 unit of nuclease *Bal*31 (Bethesda Research Laboratories) was added and allowed to react at 30°C for 10 min. The digestion was terminated by adding 13 μl 200 mM-ethyleneglycol-bis(β-aminoethyl ether)*N,N'*-tetraacetic acid, followed by extractions with phenol and chloroform. The DNA was precipitated with ethanol and resuspended in 10 mM-Tris·HCl (pH 7·5), 10 mM-$MgCl_2$, 100 mM-NaCl, 6 mM-2-mercaptoethanol, 100 μg gelatin/ml. Each deoxynucleoside triphosphate was added to 1 mM, 1·4 units *E. coli* DNA polymerase I Klenow fragment were added, and the reaction incubated at room temperature for 15 min. The enzyme was heat-inactivated, and the reaction diluted to 200 μl which included 10 mM-dithiothreitol, 1 mM-ATP, 550 ng *Eco*RI linkers (a gift from C. K. Itakura) and T4 DNA ligase (a gift from S. Scherer). The ligation reaction proceeded at room temperature overnight, and was then used to transform *E. coli* HB101 to ampicillin resistance. Transformants were colony purified, and retested for drug resistance. Small overnight cultures were grown and rapid plasmid isolations carried out. The resulting plasmid DNAs were digested with *Eco*RI and fractionated on a 1·2% agarose gel. The clone designated aDm2023Δ23 had about 950 base-pairs of *Drosophila* DNA, centered on the original *Xho*I site, removed. A large-scale preparation of this plasmid DNA was used for sequence determination.

3. Results

(a) *Three ways the 68C cluster is not regulated*

It is clear from the transcription map of the 68C gene cluster that the three mRNAs cannot derive from a common precursor, and thus that the tight co-ordination of expression of these RNAs does not result from their sharing a single promoter (Meyerowitz & Hogness, 1982).

A second way in which co-ordinate regulation of clustered genes can be accomplished is by amplification of the chromosomal region containing the genes at the time of their expression (Spradling & Mahowald, 1980; Spradling, 1981). A third mechanism by which co-ordinate expression might be triggered is by a DNA rearrangement in the chromosomal DNA of the expressing tissue (Brack *et al.*, 1978; Zieg *et al.*, 1978; Seidman *et al.*, 1979). To determine if the 68C glue gene cluster has undergone differential amplification or DNA rearrangement in the tissue of its expression in third instar larvae, high molecular weight DNA was isolated from adult flies and from third instar larval salivary glands. Equivalent amounts of each DNA preparation were digested with the restriction endonucleases *Eco*RI, *Hin*dIII and *Sal*I, subjected to electrophoresis in an agarose gel, and blotted to a nitrocellulose filter (Southern, 1975). The filter was hybridized with ^{32}P-labeled λaDm1501-10 DNA (Meyerowitz & Hogness, 1982); this phage contains 18·2 kb† of contiguous genomic DNA, including the 68C glue gene cluster (Fig. 1). The pattern of restriction fragment sizes and the autoradiographic intensities of hybridization of each fragment were identical in the adult and salivary gland lanes (Fig. 2); and the sizes of restriction fragments

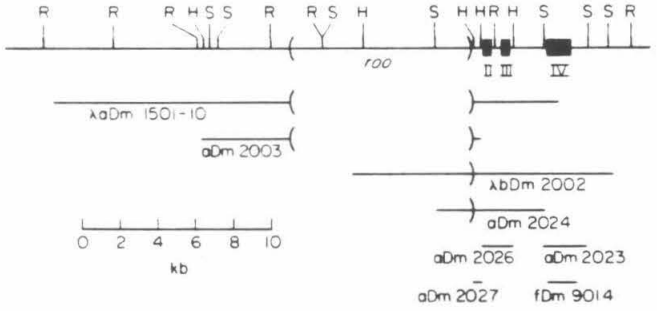


FIG. 1. λ clones and plasmid subclones used in this study. A restriction map of the studied region of the 68C puff is given, showing the position of the *roo* transposable element found in the OR16f chromosome. Genes II, III and IV are indicated as filled boxes. Cleavage sites for the restriction enzymes *Eco*RI (R), *Hin*dIII (H) and *Sal*I (S) are shown. Extents of the genomic DNA clones are indicated. λaDm1501-10 and aDm2003 have been described (Meyerowitz & Hogness, 1982). aDm2023 contains the 2·4 kb *Sal*I fragment homologous to the *Sgs-3* gene inserted into the pBR322 *Sal*I site. aDm2024 contains the 5·7 kb *Sal*I fragment adjacent to that cloned in aDm2023. aDm2026 contains the 1·65 kb *Hin*dIII fragment that includes *Sgs-7* and *Sgs-8* inserted into the pBR322 *Hin*dIII site. aDm2027 contains the 0·53 kb *Hin*dIII fragment adjacent to that present in aDm2026. fDm9014 contains the 1·6 kb *Pru*I fragment that includes *Sgs-3* inserted into the *Pru*I site of pBR325. These 5 plasmid subclones were prepared from λbDm2002 (Meyerowitz & Hogness, 1982) by routine subcloning procedures.

† Abbreviations used: kb, 10^3 base-pairs; cDNA, complementary DNA.

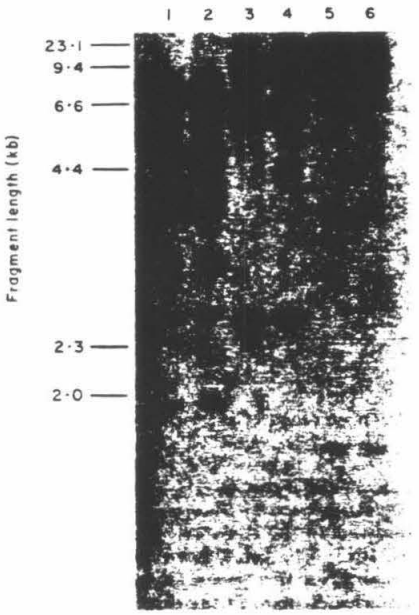


FIG. 2. 68C cluster region DNA is neither amplified nor rearranged in third instar larval salivary glands. High molecular weight DNAs from *D. melanogaster* strain OR16f adult flies (lanes 1, 3 and 5) or third instar larval salivary glands from the same strain (lanes 2, 4 and 6) were digested with *Eco*RI (lanes 1 and 2), *Sal*I (lanes 3 and 4), or *Hin*dIII (lanes 5 and 6). The digested DNAs were subjected to electrophoresis through a 0·9% agarose gel, blotted to nitrocellulose, and hybridized with ^{32}P-labeled λaDm1501-10 DNA. The size standard used was λc1857 *S*7 DNA digested with *Hin*dIII. Autoradiographs of the genome blot filter were scanned with a Joyce–Loebl microdensitometer. Peaks from the *Eco*RI digest lanes were cut out and weighed. The ratios of adult fly DNA peaks to the corresponding salivary gland DNA peaks ranged from 0·9 to 1·1. To control for the preferential polytenization of euchromatic DNA in the salivary gland, the genome blot filter was washed and rehybridized with ^{32}P-labeled aDm2040 DNA (E. M. Meyerowitz, unpublished results). This genomic clone derives from near 68C 10-11, is unique in the genome and is not detectably transcribed in third instar larval salivary glands. Microdensitometry of the resulting autoradiograph gave a ratio of adult fly DNA to salivary gland DNA of 0·8.

found were the same as those measured in clones such as λaDm1501-10 that are derived from embryonic DNA. The same filter was washed of ^{32}P-labeled probe and sequentially rehybridized with two other ^{32}P-labeled probes: aDm2026, a plasmid containing the 1·65 kb *Hin*dIII fragment that includes the coding DNA for the two small, divergently transcribed RNAs II and III, and aDm2023, a plasmid with the 2·4 kb *Sal*I fragment that contains the coding DNA for the large RNA IV (Fig. 1). These hybridizations also showed no differences between adult and salivary gland DNA, either in restriction fragment size or extent of probe hybridization.

(b) *DNA sequence of the 68C cluster region*

One possible mechanism for the co-ordinate regulation of the three 68C RNAs is that each RNA has its own equivalent of a bacterial operator, and that the

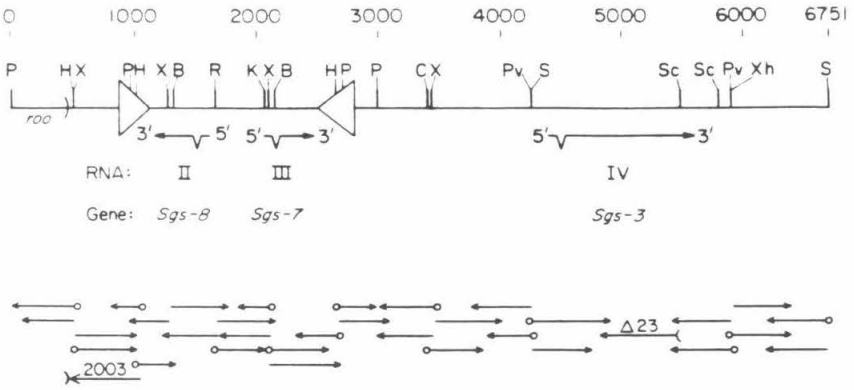


FIG. 3. Sequence determination strategy. Restriction enzyme cleavage sites are abbreviated as follows: B, *Bgl*II; C, *Cla*I; H, *Hin*dIII; K, *Kpn*I; P, *Pst*I; Pv, *Pru*I; R, *Eco*RI; S, *Sal*I; Sc, *Sac*I; X, *Xba*I; Xh, *Xho*I. The two triangles indicate the extents of the inverted repeat elements. Close-parenthesis marks the right-hand boundary of the *roo* transposable element (Meyerowitz & Hogness, 1982) inserted at 68C in the fly stock used in these experiments. Below the map are indicated the directions and extent of transcription of the three 68C genes expressed in third instar larval salivary glands (Meyerowitz & Hogness, 1982). The small downward-pointing carats indicate the locations of the intervening sequences present in the genes. Below this is a summary of the sequence determination experiments. Arrowheads point in the directions sequence data were read. The base of each arrow is aligned with the labeled restriction site. Plain arrows are sequences read 3′ to 5′ using DNA polymerase I Klenow fragment-labeled DNA. Circles represent restriction sites labeled with T4 polynucleotide kinase; these sequences were read 5′ to 3′. The arrow labeled 2003 indicates a sequence determined from the genomic subclone aDm2003 used to establish the point of insertion of the *roo* transposable element. The arrow labeled Δ23 represents sequence data obtained from the deletion derivative aDm2023Δ23 (see Materials and Methods). The *Xho*I site was read across by sequence determination from a nearby *Hae*III site. The left *Sal*I site was read across from the adjacent *Pru*I site. All restriction sites used as origins of sequence data were thus read across. Determinations were done on both strands everywhere except the leftmost 70 nucleotides of the *roo* transposable element. The scale is in base-pairs.

three operators are identical, or nearly so, and therefore respond identically to the cellular signals that control puff transcription. To seek such DNA regions, the sequence of 6751 contiguous nucleotides of the *D. melanogaster* genomic DNA clone λbDm2002, containing the coding sequences of all three of the 68C glue RNAs, was determined. Figure 3 shows a restriction map of the sequenced DNA, the relative positions of the three glue RNAs in it, and the sequencing strategy used. All but the leftmost 70 nucleotides of the sequence were determined on each of the complementary DNA strands, and complementarity was observed everywhere, although at 11 *Eco*RII sites (C-C-$\frac{A}{T}$-G-G), in which methylation occurs at the inner cytosine residue in *dcm*$^+$ *E. coli*, there was a gap rather than a band in the C and C+T lanes of the sequencing gel (Ohmori *et al.*, 1978). Agreement between the sequence and the experimentally determined sites of digestion by 11 hexanucleotide-recognizing restriction endonucleases was perfect, except that three *Cla*I sites indicated by the sequence were not cleaved by this enzyme. Each of the resistant sites overlaps the sequence G-A-T-C, where the

*Cla*I site is A-T-C-G-A-T. G-A-T-C is a site for adenine methylation mediated by the *E. coli dam*$^+$ methylase (Geier & Modrich, 1979). Methylation of adenine in the *Cla*I recognition site appears to be sufficient to prevent the restriction activity of the enzyme, as predicted by Backman (1980). The 68C nucleotide sequence is presented in Figure 4. Analysis of chromosomal rearrangements with breakpoints near the sequenced DNA indicates that the sequence is presented in telomeric to centromeric order (E. M. Meyerowitz & M. A. Crosby, unpublished work).

The beginning of the sequence, nucleotides 1 through 463, is a sequence of a part of a transposable element of the repetitive *roo* family, that is present in our *D. melanogaster* Oregon-R strain just to the left of the glue-coding gene cluster (Meyerowitz & Hogness, 1982). That these nucleotides are in the transposable element was determined by comparison of the sequence of a clone from our 68C *roo*-containing strain (aDm2024, Fig. 1) to a homologous genomic clone (aDm2003, Fig. 1, sequence not presented) from a wild-type chromosome which does not have a *roo* element adjacent to the 68C glue puff. The *roo* sequence presented here differs in only five positions from the similar sequence determined by Scherer *et al.* (1982) for their *B104* transposable element family which, as demonstrated in their paper, is equivalent to the *roo* family.

The nucleotides in positions 464 through 874 are DNA that is unique to the 68C puff region. Following this is one element of an inverted repeat sequence, from nucleotides 875 to 1159. The complementary element extends from positions 2853 to 2569. This pair of elements has been observed as a stem and loop structure in electron microscopic analysis of melted and reannealed DNA from λbDm2002 (Meyerowitz & Hogness, 1982); the elements flank the 3′ ends of the group II (sgs-8) and group III (sgs-7) RNAs. The sequence shows that each element is 285 base-pairs in length, and that the two elements are complementary at 93% (266) of their nucleotide positions (Fig. 5). The boundaries of the elements are distinct. Since no similar element appears adjacent to the 3′ end of the third 68C RNA, the group IV sgs-3 RNA, it seems unlikely that this sequence is responsible for the co-ordinate expression of all three 68C RNAs. This conclusion is supported by further observations on the inverted repeat elements: when DNA containing one or both of them is ^{32}P-labeled and used to probe plaque filters or colony filters containing DNA of λ or cosmid libraries of *D. melanogaster* genomic fragments, clones containing one region of the fly genome in addition to the 68C puff are obtained. This region contains three additional copies of the repeat element in 6 kb of contiguous sequence. The three copies are direct repeats, all in the same relative orientation. When clones containing this region are ^{3}H-labeled by nick-translation and hybridized to salivary gland polytene chromosomes (Pardue *et al.*, 1970), autoradiography shows the origin of the three additional elements to be in the 68C region, but clearly proximal to the 68C 3 to 5 position of the glue puff, and in a chromosomal area that is not puffed when the glue puff is present. ^{32}P-labeled probes containing the three proximal elements and the adjacent sequences do not give detectable signals when hybridized to RNA gel blots containing third larval instar salivary gland RNA. Likewise, ^{32}P-labeled cDNA made from third instar salivary gland polyadenylated RNA does not hybridize

strongly to the DNA of the clones containing these elements in DNA gel blot experiments. Thus, not only does one 68C glue gene lack an adjacent repeat element, but three repeat elements exist in a chromosomal region that is unpuffed in third instar salivary glands, and that does not contain coding sequences for any abundantly expressed third instar salivary gland RNA species.

The sequenced DNA between the two puff inverted repeat elements contains the coding DNA for the divergently transcribed sgs-8 and sgs-7 protein mRNAs. The region to the right of the repeat pair includes sequences 5′ to, including, and 3′ to the third puff mRNA, that coding for the sgs-3 protein.

(c) *Localization of RNA-coding DNA sequences*

To analyze the functional relevance of the sequences found in the glue gene cluster, the 5′ and 3′ limits of mRNA coding regions, and the exact positions of

FIG. 4. The complete DNA sequence of the 68C cluster region. The sequence is written 5′ to 3′ left to right. The strand shown is the template for gene II transcription (positions from approximately 1645 to 1218), and is congruent with RNA III (approx. positions 2112 to 2498) and RNA IV (approx. positions 4457 to 5646). Protein-coding nucleotides are capitalized. The restriction sites shown are identical to the map in Fig. 3, with 3 additions: *Msp*I at 2399, *Hha*I at 4726, and *Hae*III at 4835. The *roo* transposable element begins at position 463 and continues leftward. The inverted repeat elements occupy positions 875 to 1159 and 2853 to 2569. Sequence hyphens have been omitted for clarity.

any intervening sequences found within the RNAs were determined. These experiments began with the sequencing of cDNA clones representing each of the 68C puff mRNAs (the group II, group III and group IV cDNA clones of Meyerowitz & Hogness, 1982). These sequences are presented in Figure 6. Nuclease protection experiments (Berk & Sharp, 1977,1978) were then performed on the group II (sgs-8) and group III (sgs-7) mRNAs, to confirm and extend the sequence observations; while primer extension experiments (Ghosh *et al.*, 1980) were performed with the group IV, sgs-3 mRNA for the same purpose. The results for each of the mRNAs are presented in turn, starting with the leftmost.

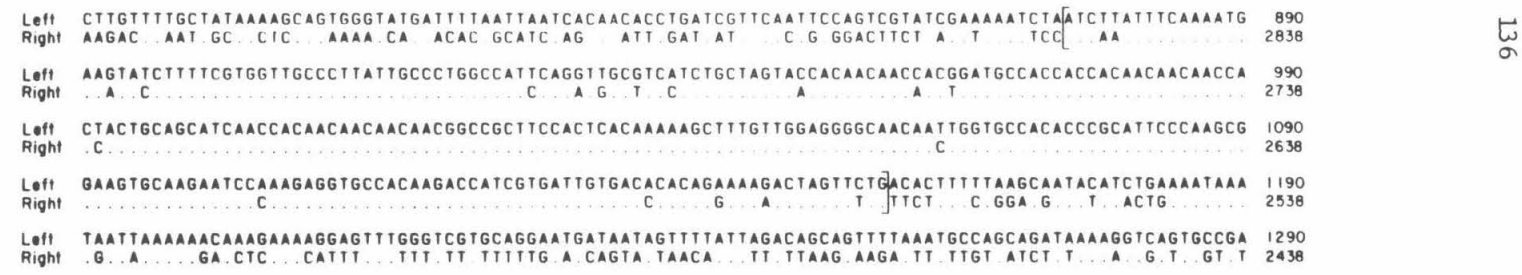


FIG. 5. Comparison of the inverted repeat elements. The sequence labeled left is positions 791 to 1290, inclusive. The sequence labeled right is complementary to positions 2937 to 2438, inclusive. Both sequences are written 5′ to 3′. Only nucleotides different from the left element are indicated for the right element. Brackets indicate the boundaries of the repeated elements. Sequence hyphens have been omitted for clarity.

Group II cDNA clone

TGAAGCTGCTCGTTGTCGCCGTCATTG|CGTGCATCATGCTCATCGGATTCGCCGATCCTGCCTCGGGCTGCAAGGATTGTTCATGCGTGATTTGTGGACC

TGGTGGCGAGCCGTGTCCTGGGTGTTCCGCACGGGTTCCCGTCTGCAAAGATCTGATCAACATTATGGTGGGTCTTGAGCGGCAGGTGCGTCAGTGCGCC

TGCGGGGAGCAGGTTTGGCTGTTCTAGAGATGTGCCCTCAACCTAATCGGCACTGACCTTTTATCTGCTGGCCTTTAAAACTGCTGTCTAATAAAACTAT

TATCATTCCTGCACGACCCA$_{31}$

Group III cDNA clone

CACCATCATCG|CTTGCATCCTGCTCATTGGATTCTCCGATCTAGCCTTGGGTGGTACCTGTGAGTACCAACCGTGTGGTCCTGGTGGAAAGGCCTGCACG

GGCTGTCCCGAAAAGCCCCAACTTTGTCAGCAGCTCATTAGCGATATTCGCAATCTCCAGCAGAAGATCCGGAAATGCGTCTGCGGAGAACCACAATGGA

TGATTTAGACACCAATCACTTTTAAAGATCACAAAAATTCTTCCTTAATAAAATTGTTATTACTGCTTC

Group IV cDNA clone

TTTGTTTTTGTCATCATCAATTGATTCTACGTGTAAGAATTAATAAAATTAGTTAGACTGCATA$_{18}$

FIG. 6. Nucleotide sequences of cDNA clones homologous to the 68C genes. Each sequence is written 5′ to 3′ left to right and represents the RNA strand. Oligo(dG) and oligo(dC) joints created during cloning are not shown. For clones II and III, the locations of the intervening sequences are indicated by vertical lines. For clones II and IV, 3′ polydeoxyadenylate is indicated by a subscript. Single base differences between the cDNA sequences and the corresponding genomic sequences are indicated by asterisks. The cDNA clones were isolated from a *Drosophila* strain heterozygous at the 68C locus (Meyerowitz & Hogness, 1982). The differences between the cDNA and genomic sequences are thus possibly genetic polymorphisms, though they may also represent errors in reverse transcription.

(i) *sgs-8*

The DNA sequence of the group II cDNA clone is homologous with the genomic DNA sequence from positions 1215 at the 3′ end of the RNA transcript, to position 1605 at the 5′ end. Genome nucleotides 1510 through 1578 are absent from the cDNA clone, indicating a 69 nucleotide intervening sequence near the 5′ end of the RNA transcript. The 3′ end of the gene can only be localized to bases 1215 to 1218; the cDNA sequence is polydeoxyadenylate beginning at 1217 and the genomic positions 1215, 1216 and 1217 are A residues. Thus, RNA termination and poly(A) addition could occur after any of positions 1218 through 1215.

To confirm the presence of an intervening sequence in the sgs-8 mRNA, and to establish an approximate location for the 5′ start of transcription, nuclease protection experiments were performed. The 404 base-pair *Xba*I-*Eco*RI fragment that includes nucleotides 1310 through 1713, and that extends from the middle of the *Sgs-8* gene to upstream of the 5′ end, was labeled at the *Xba*I site on the strand complementary to the sgs-8 mRNA using [γ-^{32}P]ATP and T4 polynucleotide kinase, then annealed to total poly(A)$^+$ RNA from third instar larval salivary glands. The hybrid was treated with either the single strand-specific nuclease S_1, or with *E. coli* exonuclease VII (Fig. 7). After nuclease S_1 digestion, the labeled fragments that remained extended to positions 1510 to 1513, as determined by polyacrylamide gel electrophoresis adjacent to size standards made by performing sequencing reactions on the intact labeled *Xba*I-*Eco*RI fragment. After correcting for the different 3′-terminal moieties in the reaction products in the experimental and size-standard lanes (Sollner-Webb & Reeder, 1979), the 3′ end of one of the major nuclease S_1 truncation products was seen to coincide with position 1510, thus confirming the presence of an intervening sequence. Exonuclease VII digests single-stranded DNA processively from 5′ and 3′ termini (Chase & Richardson, 1974). Treatment of the RNA–DNA hybrids with this nuclease reduces the labeled DNA fragment to lengths of 336, 337 and 338 nucleotides, the 3′ ends of these fragments aligning with genomic sequence positions 1645 to 1647. Since exonuclease VII leaves undigested approximately five unpaired nucleotides extending from RNA–DNA hybrids (Donahue *et al.*, 1982; Contreras *et al.*, 1982) the 5′ terminus of the sgs-8 mRNA is very near position 1640. Controls for the nuclease digestion experiments included omitting salivary gland RNA from the hybridization reaction, in which case no labeled DNA fragment was protected from digestion, and omitting nuclease treatments, in which case the 404 base-pair starting DNA fragment was recovered intact.

(ii) *sgs-7*

The DNA sequence of the group III cDNA clone includes nucleotide positions 2164 to 2498 in the genomic clone sequence, with positions 2175 to 2240 missing. The absent sequence indicates that the sgs-7 mRNA contains a 66 nucleotide intervening sequence near its 5′ end. There is no poly(A) tract at the end of the cDNA insert that represents the 3′ end of the mRNA, but the genomic sequence from positions 2499 to 2516 is a tract of 18 consecutive A residues. It thus seems

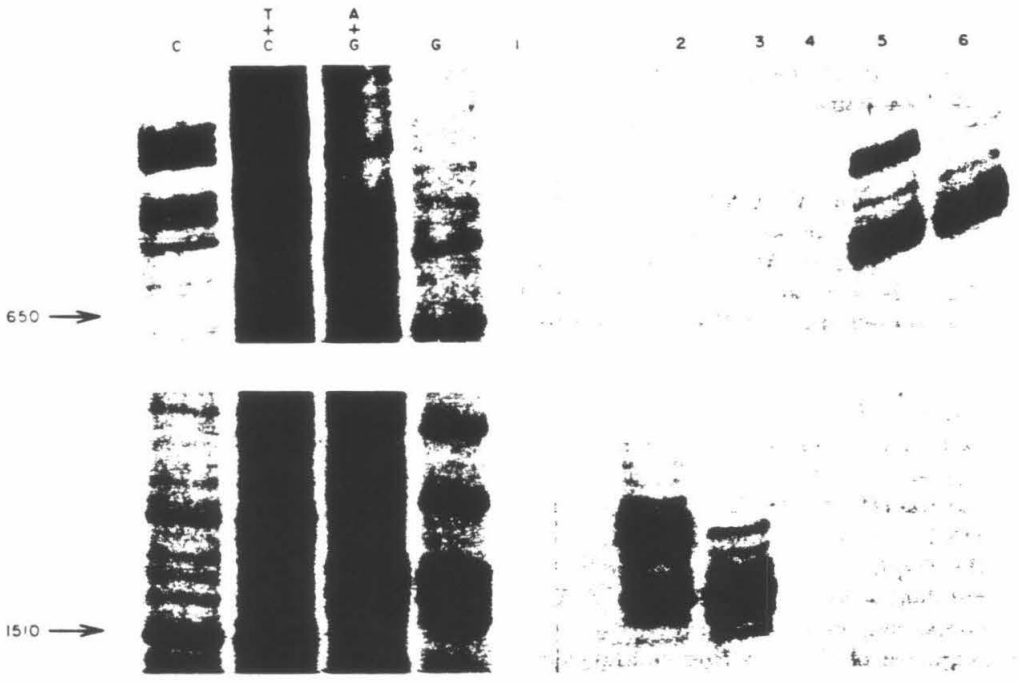


FIG. 7. Mapping of the gene II intervening sequence and 5′ end. aDm2026 DNA was *Xba*I-cut, dephosphorylated, and labeled with [γ-^{32}P]ATP and T4 polynucleotide kinase. After *Eco*RI digestion and polyacrylamide gel electrophoresis, the purified 404 base-pair *Xba*I-*Eco*RI fragment (positions 1310 to 1713) was hybridized to third instar larval salivary gland poly(A)$^+$ RNA (lanes 1 to 6) or mock hybridized to yeast tRNA. Portions of the hybridization mixes were diluted into nuclease S_1 assay buffer (lanes 1 to 3) or into exonuclease VII buffer (lanes 4 to 6). Mock digestions without added enzyme: lanes 1 and 4. Nuclease S_1 at 350 units/ml: lane 2. Nuclease S_1 at 680 units/ml: lane 3. Exonuclease VII at 4·4 units/ml: lane 5. Exonuclease VII at 8·8 units/ml: lane 6. Portions of the 404 base-pair *Xba*I-*Eco*RI fragment were subjected to sequence reactions to make size standards (lanes C, C+T, G+A, G). The fragments were separated by electrophoresis through 0·36 mm-thick 5% polyacrylamide/50% urea gels and autoradiographed. This Figure is a composite of 2 gels that were run for different times in order to resolve both sets of protected fragments. The numbers at the left show the nucleotide positions of the bands indicated.

likely that the polyadenylation site of the RNA is coded between bases 2498 and 2516 of the DNA sequence, and the possibility exists that up to 18 residues of the poly(A) tail on this RNA are added transcriptionally, rather than post-transcriptionally.

Confirmation of the intervening sequence position, and establishment of the 5′ end of the RNA were done in experiments similar to those used for the sgs-8 RNA. In this case a 683 base-pair *Msp*I-*Eco*RI fragment (positions 2400 to 1718), labeled at the *Msp*I site by use of [γ-^{32}P]ATP and T4 polynucleotide kinase, was used in hybridization to poly(A)$^+$ third instar larval salivary gland RNA. The results of nuclease S_1 digestion experiments confirmed the intervening sequence location derived from the cDNA clone sequence; the exonuclease VII

experiments placed the nucleotide coding for the 5′ end of the RNA very near the *Kpn*I site at position 2112. It is worth noting that all of the sequences between the 5′ ends of the divergently transcribed *Sgs-8* and *Sgs-7* genes total less than 500 base-pairs.

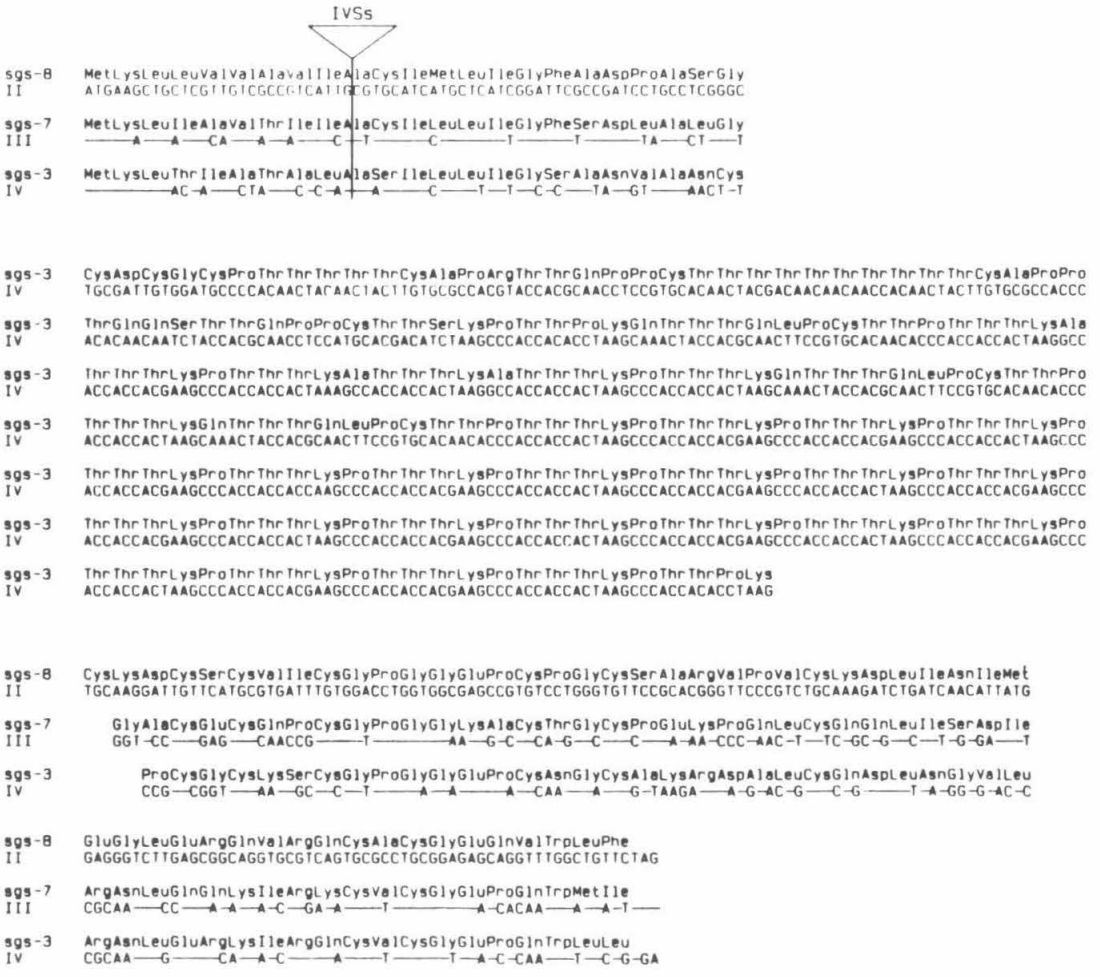
(iii) *sgs-3*

The group IV cDNA clone, representing the sgs-3 RNA, was only partially sequenced. The sequence of the end derived from the 3′ terminus of the RNA contained a poly(A) tract adjacent to genomic nucleotide position 5646, indicating that this is the poly(A) addition site of the RNA. The 5′ end of the RNA is not represented in the cDNA clone. Nuclease protection experiments performed by K. Burtis & D. Hogness (personal communication) suggested that a small intervening sequence exists in the sgs-3 RNA near position 4550. To confirm this result, and to determine the precise size and location of this sequence, a primer extension experiment was done. An 111 base-pair *Hae*III-*Hha*I genomic DNA fragment, containing nucleotide positions 4725 through 4835 and ^{32}P-labeled at the *Hae*III site (4835) using T4 polynucleotide kinase, was prepared. This labeled DNA was melted and annealed to a total third instar salivary gland RNA, and the poly(A)$^+$ RNA–primer DNA hybrids collected by oligo(dT)-cellulose column chromatography. The primed RNA was then incubated in a reaction mixture containing deoxynucleoside triphosphates and reverse transcriptase, and the sequence of the resulting end-labeled cDNA determined. The sequence showed, first, that there is an intervening sequence present in genomic DNA but absent from sgs-3 mRNA, including 73 genomic nucleotides (positions 4514 to 4586). In addition, the strongest site of primer extension termination was genomic nucleotide position 4457, indicating that the 5′-terminal nucleotide of the mRNA is coded at or near this position.

(d) *Search for potential co-ordinate control sequences*

With the positions of mRNA coding sequences established, a search for duplicated sequences found in the same position relative to each of the RNAs was made. The only such sequences are found near and including the 5′ ends of, and extending into the three genes. They are shown in Figure 8. Three alignment points are used in the Figure: the intervening sequence 5′ boundary, the 3′ boundary of this sequence, and sequences near the 5′ termini of the three RNA coding regions. In the vicinity of the transcription initiation region of each of the three genes is a conserved oligonucleotide $\begin{matrix}C\\T\end{matrix}$-A-T-C-$\begin{matrix}T\\A\end{matrix}$-G-$\begin{matrix}G\\T\end{matrix}$ which has been observed at the 5′ ends of other *Drosophila* genes (Snyder *et al.*, 1982). Approximately 30 nucleotides upstream of the 5′ end of each gene is a T-A-T-A sequence (Goldberg, 1979), which has been shown to be required for correct initiation site selection by RNA polymerase II in other eukaryotic genes (Grosschedl & Birnstiel, 1980; Benoist & Chambon, 1981; McKnight *et al.*, 1981). In addition, the *Sgs-8* and *Sgs-7* gene regions share homology at two upstream locations, underlined in Figure 8, and extending to almost 100 base-pairs 5′ of the

FIG. 8. Comparison of the DNA sequences flanking the intervening sequences and 5′ ends of the 68C genes. The 5′ termini have been aligned by the common oligonucleotide observed around other *Drosophila* initiation points (Snyder *et al.*, 1982). Exonuclease VII protection endpoints are marked by downward pointing arrows, cDNA extension endpoints are marked by upward pointing arrows. The T-A-T-A box sequences are marked by overlines and Gs. Highly conserved nucleotides flanking the splicing sites are boxed. Additional vertical lines mark the exon-intervening sequence (IVS) and intervening sequence/exon boundaries. Translation initiation codons are marked Met. Sequence homologies upstream of genes II and III are underlined. Sequence hyphens have been omitted for clarity.


IVSs
sgs-8
II
sgs-7
III
sgs-3
IV
MetLysLeuLeuValValAlaValIleAlaCysIleMetLeuIleGlyPheAlaAspProAlaSerGly
ATGAAGCTGCTCGTTGTCGCCGTCATTGCGTGCATCATGCTCATCGGATTCGCCGATCCTGCCTCGGGC
MetLysLeuIleAlaValThrIleIleAlaCysIleLeuLeuIleGlyPheSerAspLeuAlaLeuGly
MetLysLeuThrIleAlaThrAlaLeuAlaSerIleLeuLeuIleGlySerAlaAsnValAlaAsnCys
CysAspCysGlyCysProThrThrThrThrThrCysAlaProArgThrThrGlnProProCysThrThrThrThrThrThrThrThrThrCysAlaProPro
TGCGATTGTGGATGCCCCACAACTACAACTACTTGTGCGCCACGTACCACGCAACCTCCGTGCACAACTACGACAACAACAACCACAACTACTTGTGCGCCACCC
ThrGlnGlnSerThrThrGlnProProCysThrThrSerLysProThrThrProLysGlnThrThrThrGlnLeuProCysThrThrProThrThrThrLysAla
ACACAACAATCTACCACGCAACCTCCATGCACGACATCTAAGCCCACCACACCTAAGCAAACTACCACGCAACTTCCGTGCACAACACCCACCACCACTAAGGCC
ThrThrThrLysProThrThrThrLysAlaThrThrThrLysAlaThrThrThrLysProThrThrThrLysGlnThrThrThrGlnLeuProCysThrThrPro
ACCACCACGAAGCCCACCACCACTAAAGCCACCACCACTAAGGCCACCACCACTAAGCCCACCACCACTAAGCAAACTACCACGCAACTTCCGTGCACAACACCC
ThrThrThrLysGlnThrThrThrGlnLeuProCysThrThrProThrThrThrLysProThrThrThrLysProThrThrThrLysProThrThrThrLysPro
ACCACCACTAAGCAAACTACCACGCAACTTCCGTGCACAACACCCACCACCACTAAGCCCACCACCACGAAGCCCACCACCACGAAGCCCACCACCACTAAGCCC
ThrThrThrLysProThrThrThrLysProThrThrThrLysProThrThrThrLysProThrThrThrLysProThrThrThrLysProThrThrThrLysPro
ACCACCACGAAGCCCACCACCACCAAGCCCACCACCACGAAGCCCACCACCACTAAGCCCACCACCACGAAGCCCACCACCACTAAGCCCACCACCACGAAGCCC
ThrThrThrLysProThrThrThrLysProThrThrThrLysProThrThrThrLysProThrThrThrLysProThrThrThrLysProThrThrThrLysPro
ACCACCACGAAGCCCACCACCACTAAGCCCACCACCACGAAGCCCACCACCACTAAGCCCACCACCACGAAGCCCACCACCACTAAGCCCACCACCACGAAGCCC
ThrThrThrLysProThrThrThrLysProThrThrThrLysProThrThrThrLysProThrThrProLys
ACCACCACTAAGCCCACCACCACGAAGCCCACCACCACGAAGCCCACCACCACTAAGCCCACCACACCTAAG
CysLysAspCysSerCysValIleCysGlyProGlyGlyGluProCysProGlyCysSerAlaArgValProValCysLysAspLeuIleAsnIleMet
TGCAAGGATTGTTCATGCGTGATTTGTGGACCTGGTGGCGAGCCGTGTCCTGGGTGTTCCGCACGGGTTCCCGTCTGCAAAGATCTGATCAACATTATG
GlyAlaCysGluCysGlnProCysGlyProGlyGlyLysAlaCysThrGlyCysProGluLysProGlnLeuCysGlnGlnLeuIleSerAspIle
GlyProCysGlyCysLysSerCysGlyProGlyGlyGluProCysAsnGlyCysAlaLysArgAspAlaLeuCysGlnAspLeuAsnGlyValLeu
GluGlyLeuGluArgGlnValArgGlnCysAlaCysGlyGluGlnValTrpLeuPhe
GAGGGTCTTGAGCGGCAGGTGCGTCAGTGCGCCTGCGGAGAGCAGGTTTGGCTGTTCTAG
ArgAsnLeuGlnGlnLysIleArgLysCysValCysGlyGluProGlnTrpMetIle
CGCAA
ArgAsnLeuGluArgLysIleArgGlnCysValCysGlyGluProGlnTrpLeuLeu
CGCAA

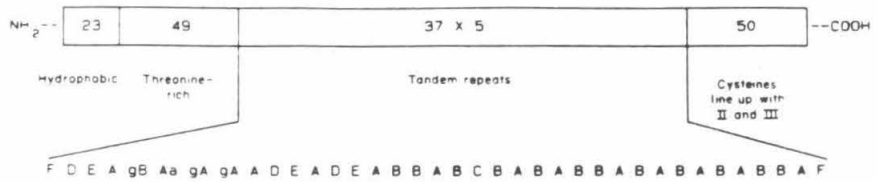
transcription initiation points. The *Sgs-3* gene region does not have either of these sequences. Within the RNA coding regions of the three genes there are considerable homologies. The 5′ untranslated regions are all similar in their DNA sequence, as are the nucleotides flanking the translation initiation codons, and those coding for the first ten amino acids. The consensus splicing donor sequences (Lerner *et al.*, 1980; Sharp, 1981) following the first ten codons are also homologous. The intervening sequences that follow show no detectable homology until five nucleotides upstream of their 3′ ends, where the 3′ end of these sequences and the consensus splicing acceptor signals are again similar. The *Sgs-8* and *Sgs-7* genes show about 40 additional nucleotides of sequence similarity downstream of their intervening sequences; the *Sgs-3* gene is homologous to the other two for only the first 19 nucleotides of this region. The next substantial homology of contiguous nucleotides found in all three genes is in the 3′ untranslated region, where about 20 nucleotides upstream of their 3′ termini each of the genes possesses the A-A-T-A-A-A sequence implicated in polyadenylation or transcription termination in many eukaryotic sequences (Proudfoot & Brownlee, 1976). The inverted repeat elements begin 58 base-pairs downstream of the sgs-8 mRNA 3′ end, and 70 base-pairs 3′ of the sgs-7 mRNA terminus. It is clear that the *Sgs-8* and *Sgs-7* genes are partly homologous throughout much of their lengths, and that, along with some 5′ sequences and their 3′ inverted repeat elements, they comprise a large, inexact inverted repetition.

(e) *Protein products of the 68C glue genes*

The amino acid sequences of the proteins coded at the 68C glue puff have been determined from the mRNA coding sequences described above (Fig. 9). Translation of eukaryotic mRNAs usually begins at the methionine codon nearest the 5′ end (Kozak, 1978); following this codon each of the 68C mRNAs has a long open reading frame. In all three cases the intervening sequence occurs between the first and second nucleotides of codon ten. The first 23 amino acids of each reading frame contain a high proportion of hydrophobic residues; these are a signal peptide removed from the primary translation products before their secretion as glue proteins (Crowley *et al.*, 1983). The sgs-8 reading frame continues for another 52 amino acids, that of sgs-7 for 51 more residues. The sgs-3 protein contains 284 amino acids following the signal sequence, the carboxy-terminal 50 of which are similar in sequence to the secreted sgs-8 and sgs-7 proteins, which are quite similar to each other. In the region at the carboxy end of the proteins eight cysteine residues and 11 other amino acid positions are identical in all three

---

FIG. 9. Complete amino acid sequences of the predicted protein products of the 68C genes. Each amino acid sequence is written in the standard 3-letter code above the corresponding mRNA-congruent DNA strand. The sequences of genes II and III, and of their protein products sgs-8 and sgs-7, are interrupted so that the amino-terminal leaders of all 3 gene products may be aligned separately from the carboxy-terminal cysteine-rich modules. Nucleotides in *Sgs-7* and *Sgs-3* that are identical to those in *Sgs-8* are indicated by horizontal lines. The intervening sequence locations are indicated (IVS). Sequence hyphens have been omitted for clarity.



Repeat type	DNA sequence	Amino acid sequence
A	CCC ACC ACC ACT AAG	Pro Thr Thr Thr Lys
gA	GCC ACC ACC ACT AAG	Ala Thr Thr Thr Lys
Aa	CCC ACC ACC ACT AAA	Pro Thr Thr Thr Lys
B	CCC ACC ACC ACG AAG	Pro Thr Thr Thr Lys
gB	GCC ACC ACC ACG AAG	Ala Thr Thr Thr Lys
C	CCC ACC ACC ACC AAG	Pro Thr Thr Thr Lys
D	CAA ACT ACC ACG CAA	Gln Thr Thr Thr Gln
E	CTT CCG TGC ACA ACA	Leu Pro Cys Thr Thr
F	CCC ACC ACA CCT AAG	Pro Thr Thr Pro Lys

FIG. 10. *Sgs-3* contains a set of imperfect tandem repeats. The overall organization of the predicted *Sgs-3* protein product is indicated by the box diagram. The numbers in each box are the numbers of amino acids found in each segment of the protein. The central tandem repetitious region has been classified into several types of 15 nucleotide repeat units. Their order is shown under the box diagram. DNA sequences and corresponding amino acid sequences of the repeat unit types are shown.

proteins. sgs-3 is different from the two small proteins by the presence of a 234 amino acid long segment between the leader peptide and the carboxy-terminal 50 amino acids. This extra segment contains an amino-terminal 49 amino acids of threonine-rich sequence, followed by 185 amino acids that are composed entirely of 37 tandem repeats of minor variants of a five amino acid unit (Fig. 10). The basic unit is Pro-Thr-Thr-Thr-Lys. The variations on this sequence, the different 15 base-pair sequences that code for the five amino acid repeats, and the higher-order repeats made of these 15 base-pair sequences are all also shown in Figure 10.

4. Discussion

(a) *Possible regulatory sequences*

It is clear from the experiments described that the basis of co-ordinate control of the three 68C RNAs does not lie in the processing of the RNAs from a common precursor, or in DNA amplification. It is also clear that large-scale DNA rearrangements are unlikely to be associated with the co-ordinate expression of the 68C glue RNAs. DNA sequencing of the 68C puff was initiated as a means of testing the hypothesis that three co-ordinately regulated genes of similar function achieve their co-ordination through possession of identical stretches of regulatory sequences. The only sequences found to be shared by the three glue protein genes and their surrounding sequences are T-A-T-A sequences and regions of protein coding sequence within the genes. T-A-T-A regions are not candidates for specific control sequences, since they are a common feature of many eukaryotic genes transcribed by RNA polymerase II. It is possible that the highly conserved 5′

translated regions of the three genes, which include the nucleotides surrounding each intervening sequence, are involved in co-ordinating the levels of the three glue protein RNAs. In particular, mechanisms of co-ordination involving common pathways of RNA splicing mediated by proteins that recognize shared sequences can be envisioned. It seems just as likely, though, that the nucleic acid sequence conservation in this region of the RNAs is a result of selection for the amino acid sequence of the hydrophobic signal peptide. No experiments have been done to differentiate between these two hypotheses.

No sequences were found which present themselves as obvious candidates for mediators of co-ordinate transcription. There are three possible reasons for this. One is that the three glue genes are not truly co-ordinately expressed, but are regulated by different mechanisms that operate only in salivary glands, and at similar enough times that their activities have not yet been distinguished. A model in which *Sgs-8* and *Sgs-7* are controlled by nearly identical regulatory sequences, but *Sgs-3* by different sequences, would be consistent with our sequence observations, and would implicate the common upstream elements of the two small genes as regulatory sequences. The second possibility is that the three genes are controlled identically by control sequences of identical function, but that we cannot recognize these sequences. This failure of recognition could occur because control sequences of identical function do not necessarily have the same DNA sequence, because the regions are too small to appear significant to us, or because the DNA sequence that comprises the regions is not composed of contiguous nucleotides, but consists of required bases separated by other, inconsequential nucleotides. The final possible reason for our failure to find identical control sequences associated with each of the 68C glue genes is that all three genes are controlled by a single set of sequences in or near the gene cluster, that affects the transcribability of the entire region, perhaps by initiating puffing as a precondition to transcription. All of these possibilities are currently being tested in this laboratory, using DNA-mediated transformation of *Drosophila* embryos (Rubin & Spradling, 1982) to assay the function of various separate fragments of the 68C gene cluster.

(b) *Protein structure and evolution*

The sequence of the DNA coding for each of the three 68C mRNAs has allowed us to predict the amino acid sequences expected of the 68C protein products. These amino acid sequences have already enabled purification and identification of the 68C proteins, and the demonstration that they are all secreted glue polypeptides (Crowley *et al.*, 1983). The amino acid sequences show that the three 68C proteins are related to each other as a clustered gene family. The members of this family show modular construction: each member has a 23 amino acid amino-terminal portion composed largely of amino acids with hydrophobic side-chains, that is not present in the mature, secreted form of the protein. Each also has a cysteine-rich carboxy-terminal set of about 50 amino acids which show considerable sequence homology between the proteins, there being 19 positions in which all three have the same residue. The sgs-7 and sgs-8 polypeptides have only these two modules.

sgs-3 contains a third module, positioned between the other two, and consisting of 234 amino acids. A total of 128 of these are threonine, with much of the module (185 residues) consisting of tandem repeats of the sequence Pro-Thr-Thr-Thr-Lys, or of sequences slightly diverged from this canonical one. The amino-terminal module probably serves as a signal peptide for protein secretion (Crowley *et al.*, 1983). The function of the carboxy-terminal module is unknown. One function of the threonine-rich module of the sgs-3 protein may be to provide a site for attachment of sugars. sgs-3 is extensively glycosylated *in vivo* (Beckendorf & Kafatos, 1976; Korge, 1977). It is unlikely that the site of carbohydrate attachment is asparagine, since the target sequence for asparagine glycosylation *via* the dolichol phosphate pathway is Asn-X-Thr or Asn-X-Ser (Staneloni & Leloir, 1982), and neither of these sequences appears in sgs-3. This leaves serine and threonine as possible sites of sugar attachment. There are only three serine residues in the processed sgs-3 polypeptide, while there are 128 threonine residues, all in the central module. An example of a protein extensively glycosylated by virtue of sugar attachment to numerous threonine residues is an anti-freeze serum protein found in Antarctic fish (Feeney & Yeh, 1978), which is modified at almost every threonine by addition of a disaccharide to the threonine hydroxyl group.

The only other *Drosophila* glue polypeptide for which sequence information is published is sgs-4, transcribed from a locus found at 3C on the polytene chromosomes (Muskavitch & Hogness, 1982). Comparison of sgs-4 amino acid or nucleic acid sequences reveals no homology between this glue polypeptide and those coded at 68C. In one respect the sgs-4 protein is similar to sgs-3: both contain substantial regions comprised of tandem repeats of a small number of amino acids. In the case of sgs-4, the repeat unit contains the seven residues Thr-Cys-Lys-Thr-Glu-Pro-Pro. Similar periodic repeats are found in a number of proteins from different sources, including silk fibroin from the moth *Bombyx mori* (Sprague *et al.*, 1979; Gage & Manning, 1980; Manning & Gage, 1980), eggshell proteins of another moth, *Antherea polyphemus* (Jones *et al.*, 1979), and zein, the seed storage protein of maize (Geraghty *et al.*, 1981; Pedersen *et al.*, 1982).

The similarities in amino acid sequence between the three 68C gene products reflect similarities in the nucleotide sequences of the genes. Since the divergence at the DNA level is substantial enough to preclude cross-hybridization under our relatively non-stringent conditions of filter hybridization and washing, nucleotide sequencing experiments were necessary to discover the relation of the three glue genes. In fact, the three members of the 68C glue gene family are so dissimilar that the order of the gene duplication events that presumably gave rise to the family cannot be deduced. As can be seen from Table 1, sgs-8 and sgs-7 appear more closely related to each other than either is to sgs-3 when the hydrophobic amino-terminal module is considered, but sgs-7 and sgs-3 are more closely related to each other than to sgs-8 when the carboxy-terminal cysteine-rich modules are compared.

The most striking feature of the evolution of this gene family is the appearance of the threonine-rich central module in sgs-3 (or its disappearance from sgs-7 and sgs-8). The gain (or loss) of this module is not mediated by intervening sequences

TABLE 1

Nucleotide and amino acid sequence homologies in the 68C glue polypeptide genes

Pair-wise comparison	Nucleotide identities†	Amino acid identities†	Amino acid similarity‡
A. *Hydrophobic amino termini*			
sgs-8-sgs-7	53/69 (77%)	15/23 (65%)	20/23 (87%)
sgs-8-sgs-3	45/69 (65%)	11/23 (48%)	17/23 (74%)
sgs-7-sgs-3	44/69 (64%)	11/23 (48%)	17/23 (74%)
B. *Cysteine-rich carboxy termini*			
sgs-8-sgs-7	81/150 (54%)	20/50 (40%)	27/50 (54%)
sgs-8-sgs-3	88/150 (59%)	27/50 (54%)	38/50 (76%)
sgs-7-sgs-3	101/150 (67%)	28/50 (56%)	38/50 (76%)

† No gaps were introduced into either set of alignments.

‡ Amino acid similarity includes both amino acid identities and "conservative" substitutions of amino acids with functionally similar side-chains (Lehninger, 1975).

at its termini, since the intervening sequences in the 68C genes are all in the middle of the amino-terminal module. Appearance (or disappearance) of the sgs-3 central module is also not due to tandem duplication, or deletion of tandem duplications: the module is unrelated to any of the other sequences of the gene. The possible importance of appearance and disappearance of modules in the evolution of structural proteins is pointed out by the employment of modular evolution in at least one other gene family, the *A. polyphemus* egg chorion proteins (Jones *et al.*, 1979). In this instance, two modules found in the middle, and at the carboxy-terminal end of the B class of chorion proteins are also found in members of the A chorion protein class. The sequences surrounding these modules are not shared between the two protein classes. In these proteins the shared regions are partly composed of tandem repeats of oligopeptides, exactly as is the threonine-rich central region of sgs-3. Studies of the 68C puff proteins in species of *Drosophila* other than *melanogaster* may help to answer some of the questions relating to evolutionary mechanisms raised by modular evolution of the 68C glue polypeptides.

We thank Dr C. M. Rice for teaching us DNA sequencing methods, T. Hunkapiller for computer programs and instruction in their use, M. Douglas for aiding in preparing computer-generated Figures, and Dr Leroy Hood for the use of his computer facility. We also thank Dr G. Scherer, K. Burtis and H. Nick for communicating unpublished results, and Drs S. Scherer and M. Snyder for discussions and advice. Two of us (M.D.G. and R.E.P.) are predoctoral fellows of the National Science Foundation. This work was supported by grant 1 RO1 GM 28075 (awarded to E.M.M.) by the Institute of General Medical Sciences, National Institutes of Health, and by a National Reserch Service Award (1 T32 GM 07616), also from the National Institute of General Medical Sciences, National Institutes of Health.

REFERENCES

Akam, M. E., Roberts, D. B., Richards, G. P. & Ashburner, M. (1978). *Cell*, **13**, 215–225.

Ashburner, M. (1972). *Chromosoma*, **38**, 255–281.

Ashburner, M. (1973). *Develop. Biol.* **35**, 47–61.
Backman, K. (1980). *Gene*. **11**, 169–172.
Beckendorf, S. K. & Kafatos, F. C. (1976). *Cell*. **9**, 365–373.
Benoist, C. & Chambon, P. (1981). *Nature (London)*. **290**, 304–310.
Berk, A. J. & Sharp, P. A. (1977). *Cell*. **12**, 721–732.
Berk, A. J. & Sharp, P. A. (1978). *Proc. Nat. Acad. Sci., U.S.A.* **75**, 1274–1278.
Brack, C., Hirama, M., Lenhard-Schuller, R. & Tanegawa, S. (1978). *Cell*, **15**, 1–14.
Casey, J. & Davidson, N. (1977). *Nucl. Acids Res.* **4**, 1539–1552.
Chase, J. W. & Richardson, C. C. (1974). *J. Biol. Chem.* **249**, 4545–4552.
Contreras, R., Gheysen, D., Knowland, J., van de Voorde, A. & Fiers, W. (1982). *Nature (London)* **300**, 500–505.
Crowley, T. E., Bond, M. W. & Meyerowitz, E. M. (1983). *Mol. Cell. Biol.* **3**, 623–634.
Davis, R. W., Botstein, D. & Roth, J. R. (1980). *Advanced Bacterial Genetics*, Cold Spring Harbor Laboratory, Cold Spring Harbor.
Donahue, T. F., Farabaugh, P. J. & Fink, G. R. (1982). *Gene*. **18**, 47–59.
Elgin, S. C. R. & Miller, D. W. (1978). In *The Genetics and Biology of Drosophila*, *2A* (Ashburner, M. & Wright, T. R. F., eds), pp. 112–121. Academic Press, New York.
Feeney, R. E. & Yeh, Y. (1978). *Advan. Protein Chem.* **32**, 191–282.
Gage, L. P. & Manning, R. F. (1980). *J. Biol. Chem.* **255**, 9444–9450.
Geier, G. E. & Modrich, P. (1979). *J. Biol. Chem.* **254**, 1408–1413.
Geraghty, D., Peifer, M. A., Rubenstein, I. & Messing, J. (1981). *Nucl. Acids Res.* **9**, 5163–5174.
Ghosh, P. K., Reddy, V. B., Piatak, M., Lebowitz, P. & Weissman, S. M. (1980). *Methods Enzymol.* **65**, 580–595.
Goldberg, M. L. (1979). Ph.D. dissertation, Stanford University, Stanford.
Gronemeyer, H. & Pongs, O. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 2108–2112.
Grosschedl, R. & Birnsteil, M. L. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 1432–1436.
Jones, C. W., Rosenthal, N., Rodakis, G. C. & Kafatos, F. C. (1979). *Cell*, **18**, 1317–1332.
Korge, G. (1975). *Proc. Nat. Acad. Sci., U.S.A.* **72**, 4550–4554.
Korge, G. (1977). *Develop. Biol.* **58**, 339–355.
Kozak, M. (1978). *Cell*. **15**, 1109–1123.
Lane, N. J., Carter, Y. R. & Ashburner, M. (1972). *Wilhelm Roux' Archiv*. **169**, 216–238.
Lehninger, A. L. (1975). *Biochemistry*. 2nd edit., Worth Publishers, Inc., New York.
Lerner, M. R., Boyle, J. A., Mount, S. M., Wolin, S. L. & Steitz, J. A. (1980). *Nature (London)*, **283**, 220–224.
Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K. & Efstratiadis, A. (1978). *Cell*. **15**, 687–701.
Manning, R. F. & Gage, L. P. (1980). *J. Biol. Chem.* **255**, 9451–9457.
Maxam, A. M. & Gilbert, W. (1977). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 560–564.
Maxam, A. M. & Gilbert, W. (1980). *Methods Enzymol.* **65**, 499–560.
McKnight, S. L., Gavis, E. R., Kingsbury, R. & Axel, R. (1981). *Cell*, **25**, 385–398.
Meyerowitz, E. M. & Hogness, D. S. (1982). *Cell*. **28**, 165–176.
Muskavitch, M. A. T. & Hogness, D. S. (1982). *Cell*. **29**, 1041–1051.
Norgard, M. V., Emigholz, K. & Monahan, J. J. (1979). *J. Bacteriol.* **138**, 270–272.
Ohmori, H., Tomizawa, J. & Maxam, A. M. (1978). *Nucl. Acids Res.* **5**, 1479–1485.
Pardue, M. L., Gerbi, S. A., Eckhardt, R. A. & Gall, J. G. (1970). *Chromosoma*. **29**, 268–290.
Pedersen, K., Devereux, J., Wilson, D. R., Sheldon, E. & Larkins, B. A. (1982). *Cell*, **29**, 1015–1026.
Proudfoot, N. J. & Brownlee, G. G. (1976). *Nature (London)*. **263**, 211–214.
Robb, J. A. (1969). *J. Cell Biol.* **41**, 876–884.
Roychoudhury, R. & Wu, R. (1980). *Methods Enzymol.* **65**, 43–62.
Rubin, G. M. & Spradling, A. C. (1982). *Science*. **218**, 348–353.
Scherer, G., Tschudi, C., Perera, J., Delius, H. & Pirrotta, V. (1982). *J. Mol. Biol.* **157**, 435–454.

Seidman, J. G., Max, E. E. & Leder, P. (1979). *Nature (London)*, **280**, 370–375.
Sharp, P. A. (1981). *Cell*, **23**, 643–646.
Smith, D. R. & Calvo, J. M. (1980). *Nucl. Acids Res.* **8**, 2255–2274.
Snyder, M., Hunkapiller, M., Yuen, D., Silvert, D., Fristrom, J. & Davidson, N. (1982). *Cell*, **28**, 1027–1040.
Sollner-Webb, B. & Reeder, R. H. (1979). *Cell*, **18**, 485–499.
Southern, E. M. (1975). *J. Mol. Biol.* **98**, 503–517.
Spradling, A. C. (1981). *Cell*, **27**, 193–201.
Spradling, A. C. & Mahowald, A. P. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 1096–1100.
Sprague, K. U., Roth, M. B., Manning, R. F. & Gage, L. P. (1979). *Cell*, **17**, 407–413.
Staneloni, R. J. & Leloir, L. F. (1982). *CRC Crit. Rev. Biochem.* **12**, 289–326.
Velissariou, V. & Ashburner, M. (1980). *Chromosoma*, **77**, 13–27.
Velissariou, V. & Ashburner, M. (1981). *Chromosoma*, **84**, 173–185.
Zhimulev, I. F. & Kolesnikov, N. N. (1975). *Wilhelm Roux' Archiv.* **178**, 15–28.
Zieg, J., Hilmen, M. & Simon, M. (1978). *Cell*, **15**, 237–244.

Edited by I. Herskowitz