STUDIES OF THE ORGANIZATION AND EXPRESSION OF INDIVIDUAL REPETITIVE SEQUENCE FAMILIES OF THE SEA URCHIN GENOME

Thesis by

James William Posakony

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1982

(Submitted March 24, 1982)

Acknowledgements

Many people contributed to my work here at Caltech and to my enjoyment of it. My advisor, Eric Davidson, taught me from his unique perspective a great deal about early development, and provided me with a place in which to try to contribute a little to our understanding of it. Roy Britten was a constant source of stimulating questions and thoughts about evolution and the structure of the genome. From both of them I learned much about the critical interpretation of experimental results.

Jane Rigg's friendship and resourcefulness have been especially valuable to me.

I have benefited in countless ways from having as my colleagues the members of the Davidson and Britten laboratories, both past and present. Lots of ideas and techniques came from members of the Attardi, Bonner, Norman Davidson, Hood, and Maniatis laboratories as well.

I want to thank Tom Maniatis and Arg Efstratiadis for letting me learn from them about gene structure and evolution.

Having also been an undergraduate at Caltech, I have been fortunate in having long-standing associations with members of the Biology Division. In particular I want to thank Giuseppe Attardi for the wonderful opportunities he gave me early on, and for his continued interest and encouragement. I am especially grateful to Ray Owen, whose counsel and friendship for me and others have been unfailing over a span of ten years.

The road has been made a lot smoother along the way by the constant helpfulness of those in the divisional and administrative offices, particularly Geraldine Cranmer, Lody Kempees, Bernie Miller, and the Biology Division secretaries.

Finally, I want to say that I feel most grateful of all to those who helped me learn new things and introduced me to new ideas during my time here—those who were constantly willing to discuss the biology of the situation, those for whom curiosity about living systems is a major motivation.

During the period of my graduate studies I was supported by an Earle C. Anthony Graduate Fellowship and by a National Institutes of Health Predoctoral Traineeship. The preparation of this thesis was supported by the Jean Weigle Memorial Fund.

This thesis is gratefully dedicated to my wife Leila and to my parents, for their love, encouragement, and patience.

Abstract

Individual repetitive DNA sequence families of the sea urchin Strongylocentrotus purpuratus were investigated with regard to their genomic organization, the internal structure of their members, and the structural and developmental characteristics of their RNA transcripts.

Analysis by gel blot hybridization and reassociation kinetics of cloned genomic DNA fragments containing members of three specific repeat families reveals a different pattern of organization in each case. One family is organized into long regions of repeated DNA, usually containing several members of the family in a tandem or clustered arrangement. A second family exists as long repeated elements occurring only once in a local genomic region. The third family consists of short repetitive sequence elements which are generally flanked on either side by single-copy sequences.

The internal structure of eight cloned repetitive sequence elements was examined by determination of their nucleotide sequences. The lack of sequence homology among the eight elements indicates that they are representative of distinct repeat families. For the most part they consist of complex sequence internally, with a minor fraction of the length of five of the eight occupied by direct or inverse sequence repetitions. Six of the eight sequences are not translatable. Comparison of the nucleotide sequences of three different members of the same repeat family reveals that they are not simply colinear sequence variants, but that they differ in the presence and/or arrangement of small sequence subelements.

Hybridization with cloned repetitive sequence elements was used to demonstrate that the level of representation of specific repeat sequences is quantitatively similar in the egg RNA of two sea urchin species, S. purpuratus and S. franciscanus.

Egg and embryo polyadenylated RNAs bearing specific repetitive sequences were analyzed by cDNA cloning, DNA and RNA gel blot hybridization, and DNA sequencing. It was found that the two complements of a given repeat are carried

on different sets of polyadenylated transcripts, which are generally quite long (>3 kilobases, with an estimated number average length of 5-6 kilobases). Within these transcripts, specific short repetitive sequence elements are found interspersed either with single-copy sequences or with other repeat sequences. It is demonstrated by sequencing that one such repeat-containing region is not translatable. The sets of polyadenylated transcripts deriving from several individual repeat families undergo substantial quantitative and probably qualitative modulation during early sea urchin development. Analysis of specific transcripts with single-copy probes from repeat-containing cDNA clones indicates that the embryo genome is transcribed to produce at least some of the same interspersed RNAs as are stored in the oocyte during oogenesis. Finally, the transcripts bearing specific repeat sequences in the polyadenylated egg RNA of two related sea urchin species were found to be qualitatively dissimilar.

Table of Contents

Introduction

The first three chapters of this thesis are presented in the form of reproductions of published papers. Here I would like to provide a fuller explanation of their background, bring together in one place the logical connections between them, and indicate their relationship with the experiments described in Chapter IV.

Genomic Organization of Individual Repetitive Sequence Families

One issue of basic importance to us concerned the way in which the members of a given repetitive sequence family are arranged and distributed in the genome. The overall pattern of interspersion of repetitive and single copy sequences in many genomes [e.g., Xenopus (Davidson et al., 1973; Chamberlin et al., 1975); sea urchin (Graham et al., 1974); Drosophila (Manning et al., 1975); Dictyostelium (Firtel and Kindle, 1975); and human (Schmid and Deininger, 1975)] and the existence of "long" and "short" elements of repeated DNA (Goldberg et al., 1975) had been known for some time, but little specific information existed regarding the genomic organization of individual non-satellite repeat families. The experiments presented in Chapter I reveal a remarkable diversity in the way specific families are arranged. The 2034 and 2108 repeat sequences were both known to belong to the "long" class of repeats (Moore et al., 1981), but as shown in Chapter I, the members of the 2034 family occur in long clustered or tandem arrays involving several copies of the 2034 sequence, whereas the 2108 family consists of several-kilobase-long units (see also Scheller et al., 1981) which occur singly in a given genomic region. By contrast, the 2109 family is made up of "classical" short (200-300 nucleotide) repetitive elements, generally flanked on either side by single-copy sequences. This diversity of organization raises major questions concerning the ways in which different repetitive sequence families evolve and about the mechanism(s) of their dispersal in the genome. At the same time, the specific information obtained is extremely valuable in our interpretation

of the structure of RNA transcripts deriving from these families (Chapter IV).

Internal Structure of Repetitive Sequence Elements

The experiments of Klein et al. (1978) examined for the first time the characteristics of specific repetitive sequence families, by making use of individual cloned repeat elements in DNA reassociation and thermal denaturation studies. That the reiterated fraction of the genome was composed of more-or-less distinct sets (families) of homologous sequences (Britten and Kohne, 1968) was confirmed: quantitative estimates of reiteration frequency (number of family members per genome) were obtained for 26 families, and these varied from a few to several thousand. It was known from earlier studies (Britten and Kohne, 1968; Davidson et al., 1974) that the various members of a given repeat family are not precise copies of each other, so that when complementary strands from different members are reannealed, the thermal stability of these heteroduplexes is found to be lower than that of the native homoduplexes or of reassociated single-copy DNA. Thus, when individual cloned repeat elements were reacted with total genomic DNA (Klein et al., 1978), or with single isolated repeat family members (Scheller et al., 1981), the thermal stability of the resulting heteroduplexes was lower than that of the duplex probe element itself, and the magnitude of this difference varied from family to family and member to member. Both these and the earlier studies were interpreted in terms of base sequence mismatch (divergence) between individual members of a repeat family (Bonner et al., 1973), implying that within the most divergent families, more than 20% mismatch was the rule (Klein et al., 1978; Scheller et al., 1981).

The studies described in Chapter II were undertaken to examine by nucleotide sequence determination the internal structure of a variety of repetitive sequence elements. In the first part of this investigation, sequences were obtained for eight cloned repeat elements, selected from those studied by Klein et al. (1978) as being

representative (if such is possible), both in terms of reiteration frequency and degree of apparent intrafamilial sequence divergence. We were interested primarily in two questions: (1) the character (simple or complex) of the sequence making up a repeat element, and whether specific features of internal organization (direct or inverse repetitions, etc.) could be discerned; (2) whether different repetitive sequences, apparently representing different repeat families, nevertheless carry short sequence elements in common.

All eight elements were found to be composed very largely of complex sequence. Short but statistically significant direct and inverse repetitions formed a minor fraction of the length of five of the eight sequences. The general lack of statistically significant sequence homologies showed that the eight elements are distinct. These observations provide strong support for the earlier interpretations (e.g., Graham et al., 1974) of DNA reassociation kinetic data as indicating a high sequence complexity for the moderately reiterated portion of the genome. It appears very likely that other genomes with a substantial content of repeated DNA will also be found to contain a large number of distinct sequence families, despite recent claims to the contrary (e.g., see Deininger et al., 1981).

The observation by Costantini et al. (1978) and Scheller et al. (1978) that (in contrast to single-copy sequences) both complements of several individual repeat sequences are represented at approximately equal levels in sea urchin RNA's led to the suggestion that, even if repetitive elements form a part of certain messenger RNA's in the sea urchin, these elements would not play a role in protein coding. Direct evidence on this point was provided by the finding that six of the eight repeat elements which had been sequenced were translationally blocked by termination codons in all six possible reading frames. Of course, it was conceivable that the transcribed members of repeat sequence families might contain open reading frames. It was subsequently possible to show that a repetitive sequence region actually appearing

in egg polyadenylated RNA is in fact non-translatable (see Chapter IV).

In the second part of the investigation of repeat element fine structure (Chapter II), the nucleotide sequences of three different members of the same repetitive sequence family (2109) were determined. The results of Chapter I showed that this family is composed of short repeat elements which are in general embedded in single-copy DNA. The measurements of Klein et al. (1978) and Scheller et al. (1981) identified the 2109 family as having a high degree of intrafamilial sequence divergence. Now we were interested in defining exact relationships among its members, to see whether the base differences were distributed along the length of the element, or whether invariant regions could be discerned. The sequence analysis revealed that, contrary to our expectation, the members of the 2109 family are not simply colinear sequence variants, but instead differ from each other in the presence or absence and order of arrangement of small subelements. Moreover, much less base sequence divergence was found between the three elements than was predicted by the thermal stability measurements. These observations allowed two important conclusions:

(1) That a re-interpretation was required of the low thermal stability of heteroduplexes between 2109 family members. A consequence of the variable subelement composition of different 2109 repeats is that the heteroduplexes they can form are often quite short (<100 nucleotides) and this by itself has a major effect on thermal stability (Britten et al., 1974). When heteroduplex length is taken into account, together with the base sequence mismatch which is observed, the measured thermal stabilities can be satisfactorily accounted for. This strongly suggests that the 2109 family, and perhaps others, have undergone much less sequence divergence than previously supposed. Analysis of a 2109 family member found within an egg RNA cDNA clone shows that it is 94% homologous over a span of 62 nucleotides with the probe sequence CS2109A (see Chapter IV); nevertheless, a heteroduplex between them would be expected to melt 15-20°C below the t_m of either homoduplex.

(2) That fine-scale sequence rearrangement has occurred during the evolution of the 2109 repeat family. The process by which such small (<100 nucleotide) sequence segments are scrambled or lost is unclear, but it may be connected to the mechanism responsible for the dispersal of short repeat elements in the genome. In any case, it is clear that both the number (Moore et al., 1978) and the subelement composition (Chapter II) of 2109 repeat sequences are evolutionarily fluid.

Evolutionary Conservation of Repeat Sequence Representation in Egg RNA

Using the cloned repetitive sequence elements whose families in the S. purpuratus genome had been studied by Klein et al. (1978), Moore et al. (1978) investigated the characteristics of these same repeat families in the genome of S. franciscanus, a congener of S. purpuratus. It was found that striking changes in the sizes of many repetitive sequence families have occurred in the 10-20 million years (Durham, 1966) since the divergence of lineages leading to the two species. In general, the reiteration frequencies of the repeats studied were lower in S. franciscanus than in S. purpuratus by as much as 20-fold. (It may safely be assumed that, had the cloned repeat elements used as probes been derived from S. franciscanus DNA rather than from S. purpuratus, the reverse result would have been obtained, since Moore et al. provided conclusive evidence to this effect using kinetically isolated repetitive fractions from each genome.) Nevertheless, the interspecies difference in the mean thermal stability of heteroduplexes between a cloned repeat element and total genomic DNA was in all cases less than the interspecies difference in single-copy DNA (Hall et al., 1980). This suggested some degree of selective conservation of repeat sequences.

Costantini et al. (1978) carried out quantitative measurements of the overall level of representation in S. purpuratus egg RNA of many of these same repetitive sequences. Given their generally lower reiteration frequency in the genome of S. franciscanus, it was of great interest to determine their level of representation

in the egg RNA of that species. These experiments are described in Chapter III. Very simply, it was found that the concentration of transcripts complementary to the different repeat elements was very similar in the two egg RNAs, despite the differences in repeat family sizes in the two genomes. This observation led in turn to experiments described in Chapter IV, which show that polyadenylated transcripts bearing specific repeat sequences are different in size and prevalence in S. purpuratus vs. S. franciscanus egg RNA.

References

Bonner, T. I., Brenner, D. J., Neufeld, B. R., and Britten, R. J. (1973). J. Mol. Biol. **81**:123-135.

Britten, R. J., and Kohne, D. E. (1968). Science **161**:529-540.

Britten, R. J., Graham, D. E., and Neufeld, B. R. (1974). In Methods in Enzymology (Grossman, L., and Moldave, K., eds.), vol. 29E, pp. 363-406, Academic Press, New York.

Chamberlin, M. E., Britten, R. J., and Davidson, E. H. (1975). J. Mol. Biol. **96**:317-333.

Costantini, F. D., Scheller, R. H., Britten, R. J., and Davidson, E. H. (1978). Cell **15**:173-187.

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

Davidson, E. H., Graham, D. E., Neufeld, B. R., chamberlin, M. E., Amenson, C. S., Hough, B. R., and Britten, R. J. (1974). Cold Spring Harbor Symp. Quant. Biol. **38**:295-301.

Deininger, P. L., Jolly, D. J., Rubin, C. M., Friedmann, T., and Schmid, C. W. (1981). J. Mol. Biol. **151**:17-33.

Durham, J. W. (1966). In Treatise on Invertebrate Paleontology (U) Echinodermata (Moore, R. C., ed.), vol. 3, part 1, pp. 270-295, The Geological Society of America and The University of Kansas Press, New York.

Firtel, R. A., and Kindle, K. (1975). Cell **5**:401-411.

Goldberg, R. B., Crain, W. R., Ruderman, J. V., Moore, G. P., Barnett, T. R., Higgins, R. C., Gelfand, R. A., Galau, G. A., Britten, R. J., and Davidson, E. H. (1975). Chromosoma (Berl.) **51**:225-251.

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

Hall, T. J., Grula, J. W., Davidson, E. H., and Britten, R. J. (1980). J. Mol. Evol. **16**:95-110.

Klein, W. H., Thomas, T. L., Lai, C., Scheller, R. H., Britten, R. J., and Davidson, E. H. (1978). Cell **14**:889-900.

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

Moore, G. P., Scheller, R. H., Davidson, E. H., and Britten, R. J. (1978). Cell **15**:649-660.

Moore, G. P., Pearson, W. R., Davidson, E. H., and Britten, R. J. (1981). Chromosoma (Berl.) **84**:19-32.

Scheller, R. H., Costantini, F. D., Kozlowski, M. R., Britten, R. J., and Davidson, E. H. (1978). Cell **15**:189-203.

Scheller, R. H., Anderson, D. A., Posakony, J. W., McAllister, L. B., Britten, R. J., and Davidson, E. H. (1981). J. Mol. Biol. **149**:15-39.

Schmid, C. W., and Deininger, P. L. (1975). Cell **6**:345-358.

Chapter I

Reprinted from *J. Mol. Biol.* (1981) **145**, 5–28

Repetitive Sequences of the Sea Urchin Genome

Distribution of Members of Specific Repetitive Families

David M. Anderson, Richard H. Scheller, James W. Posakony
Linda B. McAllister, Steven G. Trabert, Clifford Beall
Roy J. Britten and Eric H. Davidson

J. Mol. Biol. (1981) **145**, 5–28

Repetitive Sequences of the Sea Urchin Genome

Distribution of Members of Specific Repetitive Families

DAVID M. ANDERSON, RICHARD H. SCHELLER, JAMES W. POSAKONY
LINDA B. MCALLISTER, STEVEN G. TRABERT, CLIFFORD BEALL
ROY J. BRITTEN, AND ERIC H. DAVIDSON

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

(Received 28 May 1980, and in revised form 30 August 1980)

Three repetitive sequence families from the sea urchin genome were studied, each defined by homology with a specific cloned probe one to a few hundred nucleotides long. Recombinant λ–sea urchin DNA libraries were screened with these probes, and individual recombinants were selected that include genomic members of these families. Restriction mapping, gel blot, and kinetic analyses were carried out to determine the organization of each repeat family. Sequence elements belonging to the first of the three repeat families were found to be embedded in longer repeat sequences. These repeat sequences frequently occur in small clusters. Members of the second repeat family are also found in a long repetitive sequence environment, but these repeats usually occur singly in any given region of the DNA. The sequences of the third repeat are only 200 to 300 nucleotides long, and are generally terminated by single copy DNA, though a few examples were found associated with other repeats. These three repeat sequence families constitute sets of homologous sequence elements that relate distant regions of the DNA.

1. Introduction

Renaturation kinetics indicate that animal DNAs contain a large variety of diverse repetitive sequence families. A repeat family may be defined experimentally as that set of homologous sequences which reacts with a given cloned repetitive sequence probe. Klein *et al.* (1978) identified a number of such families in previous studies on sea urchin DNA, and three of these have been chosen for the present experiments. The genome of the sea urchin displays the short period pattern of repetitive sequence interspersion typical of most groups of animals (Graham *et al.*, 1974; Davidson *et al.*, 1975). In genomes of this type, almost nothing is known about how the individual sequences belonging to repeat families are distributed with respect to each other. Members of given repeat families might be tightly clustered or they could be distributed widely in the genome. To approach this issue, we isolated from λ genome libraries a number of λ–sea urchin DNA recombinants that include genomic members of the three repeat sequence families and their flanking sequences. These recombinants were used to establish the extent to which members of the same

0022-2836/81/010005-24 $02.00/0

family occur in clustered arrays, as opposed to occurring singly in different genomic regions. In addition, we determined the sequence environment characteristic of each family; i.e. whether its members are characteristically embedded in single copy sequence or in other repetitive sequences, or both. The particular families we have chosen for this study include a set of short repetitive sequences interspersed in a typical way with single copy DNA, and two examples of long repetitive sequences (Britten *et al.*, 1976). Though the detailed pattern of organization of each family is characteristic, our measurements imply that the members of all three families occur in widely separated regions of the genome.

2. Materials and Methods

(a) *Preparation of unlabeled sea urchin DNA*

DNA was extracted from *Strongylocentrotus purpuratus* sperm, as described by Britten *et al.* (1974). Care was taken to avoid mechanical shearing of the DNA to ensure maximum double-stranded length. Driver DNA used in renaturation reactions was sheared by forcing the DNA solution through a needle valve at 50,000 lb/in^2 (Britten *et al.*, 1974). The sheared driver DNA had a weight average length of 600 nucleotides, measured by velocity sedimentation through alkaline sucrose gradients. Unsheared DNA had a length in excess of 100,000 nucleotides as measured by electron microscopy.

(b) *Partial* Eco*RI digestion of sea urchin DNA*

To prepare DNA fragments of a length suitable for cloning (15 to 20 kb†), unsheared (>100 kb) DNA was subjected to partial *Eco*RI digestion in 0·06 M-Tris·HCl (pH 8·0), 6 mM-$MgCl_2$, 0·05 M-NaCl at 37°C. We would expect an *Eco*RI recognition site to occur about once every 4000 (4^6) nucleotides (uncorrected for base composition). The *Eco*RI digestion conditions were adjusted to cleave an average of 1 site in 5 by varying the length of digestion and the ratio of enzyme to DNA. The partially cleaved DNA (200 μg) was fractionated on preparative 10% to 40% linear sucrose gradients (0·1 M-NaCl, 10 mM-Tris·HCl (pH 8·0), 10 mM-EDTA) in a Beckman SW27 rotor. Gradients were centrifuged at 24,000 revs/min for 20 h at 15°C, and 0·5 ml fractions were collected. Samples of the fractions were analyzed by electrophoresis on 0·5% agarose gels using *Eco*RI-digested Charon 4 DNA (Blattner *et al.*, 1977) as a molecular weight standard. The fractions containing DNA fragments 15 to 20 kb long were pooled and precipitated in isopropanol.

(c) *Preparation of Charon 4 vector DNA*

Charon 4 phage were grown essentially as described by F. R. Blattner in the protocol provided with the Charon λ phages. Phage were purified as described by Yamamoto *et al.* (1970). The phage DNA was extracted as follows. The purified phage were dialyzed from CsCl against 0·01 M-Tris·HCl (pH 8·0), 0·1 M-NaCl, 0·001 M-EDTA. Sodium dodecyl sulfate was added to 0·1% (w/v), and the solution was heated for 10 min at 60°C. The DNA was then extracted twice with phenol, once with chloroform/isoamyl alcohol (24:1, v/v), dialyzed against 0·1 M-Tris·HCl (pH 8·0), 0·001 M-EDTA, and precipitated with ethanol. The Charon 4 vector arms were isolated from the midpieces and prepared for use as described by Maniatis *et al.* (1978).

(d) *λ-recombinant genome libraries*

The Charon 4 arms were ligated to the partially digested sea urchin DNA at a 1·5 molar excess of vector to sea urchin DNA fragments. The ligation was carried out in 66 mM-Tris·HCl (pH 7·6), 10 mM-$MgCl_2$, 1 mM-ATP, 15 mM-dithiothreitol, 200 μg gelatin/ml for

† Abbreviations used: kb, kilobases (10^3 bases); SDS, sodium dodecyl sulfate; p.f.u., plaque-forming units.

18 h at 15°C, at a total nucleic acid concentration of 200 μg/ml. *In vitro* packaging and plate lysate amplification were carried out essentially as described by Maniatis *et al.* (1978). Cloning efficiencies of 1×10^4 to 2×10^4 plaque-forming units/μg DNA were routinely obtained. Two λ libraries constructed in this manner were utilized in this work, containing $1{\cdot}4 \times 10^5$ and $2{\cdot}5 \times 10^5$ different phage, respectively. The first of these libraries (denoted $Sp\lambda R_1A$) was derived from DNA of the sperm of 5 males and the second (denoted $Sp\lambda R_1B$) from sperm of a single male. A Poisson calculation indicates that the number of clones included in these libraries would contain 93% and 99%, respectively, of the sequences in the *S. purpuratus* genome; i.e. if there is no selection against given sea urchin DNA sequences, and if *Eco*RI sites are randomly distributed. Both of these propositions are probably untrue in detail. However, a direct complexity measurement in which DNA of the smaller library was used to drive a labeled single copy tracer showed that at least 90% of the genomic complexity was included in the amplified library at a sequence concentration ≥ 1/3 the average library single copy DNA sequence. A third and larger library used for some experiments was constructed by ligating *Eco*RI linkers onto sea urchin DNA fragments that had been partially digested with *Hae*III. This library was built and characterized by M. Chamberlin and G. Moore of this laboratory, by the methods described by Maniatis *et al.* (1978). This library (denoted $Sp\lambda H_3C$) contained $7{\cdot}8 \times 10^5$ individual phage. The average insert length in these libraries was 15 kb, and the range of length was 10 to 18 kb.

(e) *Preparation, labeling, and strand separation of cloned repetitive sequence fragments*

Recombinant plasmids containing repetitive sequence elements inserted by blunt end ligation of *Eco*RI linkers were described earlier (Scheller *et al.*, 1977). The vector was RSF2124 (So *et al.*, 1975). Briefly, sea urchin DNA was renatured to C_0t 40† and digested with S_1 nuclease, and the blunt-ended repeat duplexes that survived were cloned as indicated. Superhelical DNA of these plasmids was isolated on isopycnic CsCl gradients (Scheller *et al.*, 1977). The DNA was dialyzed into 5 mM-Tris·HCl (pH 8·0) and stored at −20°C. Plasmid DNA was cleaved with *Eco*RI to release the inserted repetitive sequence fragment and precipitated with ethanol. The DNA was dissolved, then treated with bacterial alkaline phosphatase, end-labeled with ^{32}P by the polynucleotide kinase reaction (Maxam & Gilbert, 1977), and strand separated after alkali denaturation on neutral acrylamide gels, as described by Scheller *et al.* (1978). The DNA fragments of interest were localized by autoradiography and individual bands were excised with a razor blade. Gel slices were crushed with a glass rod in sterile plastic tubes; 1 ml of 0·12 M-sodium phosphate buffer (pH 6·8), 0·05% (w/v) SDS and 10 μg of purified calf thymus DNA carrier were added to each sample, and the mixture was incubated overnight at 37°C. The acrylamide was removed by filtering through glass wool and rinsing with an additional 1 ml of phosphate buffer. The eluate contained from 80 to 95% of the radioactivity in the gel band. The sample was then incubated overnight at 60°C and passed over a 1 ml column of hydroxyapatite at 50°C. The unbound fraction contained the strand-separated repetitive sequence. The final preparations were contaminated only 0 to 4% with their complementary strands. Reactivity of these DNA preparations with excess total sea urchin DNA ranged from 55 to 98%. The non-reactive component(s) were of low molecular weight and most probably consisted of unincorporated [^{32}P]ATP.

(f) *Library screening*

The amplified sea urchin λ libraries were screened with ^{32}P-labeled cloned repeat tracers by a modification of the Benton & Davis (1977) procedure. A total of 1×10^4 to 2×10^4 recombinant phage were plated on 4×10^8 bacterial cells on 15 cm agar plates. To prevent top agar from adhering to the nitrocellulose filter when it was lifted from the plate, 0·7% (w/v) agarose rather than agar was used for plating. Phage were adsorbed to nitrocellulose filters (Schleicher and Schuell, 0·45 μm pore size) for about 10 min. The DNA was denatured and bound to the filters, as described by Benton & Davis (1977).

† C_0t, moles nucleotide liter^{-1} s.

To hybridize the filters with a labeled probe, filters were preincubated for 1 h in 4 × SET (SET is 0·15 M-NaCl, 0·03 M-Tris·HCl (pH 8·0), 2 mM-EDTA) plus 5 × Denhardt's solution (Denhardt, 1966) and 0·1% (w/v) SDS. Subsequent hybridizations were carried out in the same solution in sealed plastic bags. In general, $\sim 2 \times 10^5$ cts/min of tracer ($\sim 10^7$ cts/min per μg) were added per filter. Incubations were for 18 h at 50°C, or appropriately raised temperatures with higher salt concentration to produce an equivalent criterion condition, unless otherwise noted. Following hybridization, filters were washed several times in SET plus 0·5% (w/v) SDS at the incubation temperature. This set of conditions is described in the text as "low criterion". "High criterion" screens were carried out at 70°C, 0·1 M-NaCl, and the filters were washed in SET at 70°C. The filters were blotted dry, mounted on cardboard and exposed to preflashed Kodak XR5 X-ray film with Dupont Cronex NR Xtra Life Lightning Plus intensifying screens at −70°C for 1 to 7 days. Positive plaques from the region of a plate corresponding to a spot on the autoradiogram were picked and suspended in 1 ml of SM buffer (0·05 M-Tris·HCl (pH 7·4), 0·05 M-NaCl, 0·01 M-$MgCl_2$). The phage were titered and rescreened at a density of 200 phage per 15 cm plate. Individual positive plaques were then selected, resuspended in 1 ml of SM buffer and amplified in 2-ml liquid cultures. These were prepared by adding 10 μl of late log phase bacteria and 100 μl of the resuspended plaque to 2 ml of broth. The cultures were shaken at 37°C until lysis was evident (about 18 h). Titers of these cultures were of the order of 1×10^{10} p.f.u./ml. The lysate was cleared of debris by centrifugation and stored at 4°C over a drop of chloroform.

(g) *DNA renaturation*

DNA renaturation was carried out in 0·12 M or 0·41 M-phosphate buffer (pH 6·8) with 0·05% (w/v) SDS in sealed capillary tubes. All C_0t values quoted in the text are equivalent C_0t (Britten *et al.*, 1974). For example, C_0t in 0·41 M-phosphate buffer was converted to equivalent C_0t by multiplying by 5. Renaturation kinetic analyses carried out in this work included an internal single copy DNA rate standard. Single copy [^{3}H]DNA was prepared and labeled by gap translation (Galau *et al.*, 1976). The fraction of DNA fragments in molecules containing duplexes after reassociation was determined by binding to hydroxyapatite (DNA grade, BioRad, lot 17653). Samples were diluted to 0·12 M-phosphate buffer and applied to water-jacketed columns at 55°C. Unbound material was removed by washing with at least 5 bed volumes of 0·12 M-phosphate buffer. The duplex fraction was subsequently eluted by raising the temperature of the column to 98°C. Less than 250 μg of DNA was loaded per cm^3 of packed hydroxyapatite.

(h) *Gel blots and restriction digests*

Digestion with various restriction enzymes were carried out under the conditions suggested by the manufacturers. Transfer of DNA from agarose gels to nitrocellulose filters was as described by Southern (1975). Hybridization conditions were as described above for library screening.

3. Results

(a) *Three repetitive sequence families: general characteristics from reactions of cloned probes with genomic DNA*

The average properties of the repetitive sequence families chosen for these experiments are summarized in Table 1. Most of the data listed are quoted from other studies, and were obtained mainly by analyses of the heteroduplexes formed by reacting genomic DNA with plasmid clones bearing the three individual repeat sequences (Klein *et al.*, 1978; Moore *et al.*, 1980). The three repeat families described in Table 1 are designated according to the plasmid clones that define them; viz. clones CSp2034, CSp2108 and CSp2109.

TABLE 1

Characteristics of three repetitive sequence families

Family[a]	Length of probe sequence[b]	Repeat length class[c]	Family size: Genomic reiteration frequency[d]	Intrafamilial thermal stability[e] t_m(°C)	Representation of family in RNA[f] Oocyte RNA	Gastrula nRNA	Intestine nRNA
2034	498	Long	2500	4	0·2	0·9	48
2108	204	Long	20	6	590	95	55
2109A	180		900	≥20	0·3	1	22
2109B	110	Short	1000	≥25	5	64	1·8

[a] Repetitive sequences were isolated and cloned in the *Eco*RI site of the plasmid RSF2124. The 2109A and 2109B probes were isolated from the same plasmid, CSp2109, where they are separated by an *Eco*RI site (Posakony *et al.*, unpublished results).

[b] Lengths of the cloned repeat fragments have been obtained from their primary sequences (Posakony *et al.*, unpublished results). The values shown represent the total length (in nucleotides) of sea urchin DNA present in the plasmids CSp2034, CSp2108 and CSp2109.

[c] Both long (≥2000 nucleotides) and short (~300 nucleotides) repeat classes have been identified in the sea urchin genome (Britten *et al.*, 1976; Eden *et al.*, 1977; Moore *et al.*, 1980). Long repeats are defined as DNA sequences that are excluded from Sepharose CL2B after low C_0t incubation followed by S_1 nuclease digestion, while the short repeats are retarded in gel filtration columns to an extent indicating a mean length of about 300 nucleotides. This is also the typical length of short repeat duplexes observed in electron micrographs of partially renatured DNA (see e.g. Chamberlin *et al.*, 1975). The observed repeat length may underestimate the actual length of the repetitive sequence elements, since the renatured duplexes will terminate at the beginning of any non-homologous sequences, whether these are single copy sequences or other repetitive sequences. While most sea urchin short repeats are terminated by single copy sequence (Graham *et al.*, 1974), this cannot be taken for granted in any specific case. The 2108 and 2034 data are from Moore *et al.* (1980). 2109A and 2109B data are from these studies and from Scheller *et al.* (unpublished results).

[d] Measured by the renaturation kinetics of reactions between the labeled cloned probes and excess genomic DNA, assayed by hydroxyapatite binding. Data for 2109A and B are from Moore *et al.* (1980; and unpublished data). The conditions for which these determinations are accurate are 55°C, 0·12 M-phosphate buffer (pH 6·8). Klein *et al.* (1978) gave 1000 copies as the size of the 2034 family. However, the average of the 4 independent kinetic determinations shown later in this paper for 2034 sequences indicate that the proper value is 2500 copies per haploid genome. The measurement shown for 2108 sequences is that from Klein *et al.* (1978).

[e] Intrafamilial thermal stability (Δt_m) is the mean thermal stability of the cloned probe fragment minus the mean thermal stability of the population of heteroduplexes formed by reacting the cloned probe with genomic DNA (Klein *et al.*, 1978). The Δt_m values shown could be due to scattered mismatch or to short length of homologous sequence elements, or a combination of both factors. Mean thermal stabilities were determined (Klein *et al.*, 1978) as the temperature at which 50% of the homologous duplex or heteroduplex populations eluted from hydroxyapatite columns. If there is a large sequence divergence among the members of the family, the observed thermal stability will be affected by the criterion of incubation.

[f] Percentage representation in oocyte total RNA, gastrula nuclear RNA (nRNA) and intestine nRNA is calculated according to the equation:

$$\% \text{ representation} = \frac{T_c}{F_c \times T_{sc}} \times 100,$$

where T_c is the number of transcripts complementary to the cloned tracer per cell, T_{sc} is the number of copies of a typical single copy transcript, and F_c is the genomic reiteration frequency of the repeat family to which a given clone belongs (Scheller *et al.*, 1978).

Table 1 shows that the three repeat families differ from each other in several important characteristics. Most of the individual sequence elements belonging to the 2034 family are recovered in the "long" repeat class; i.e. they occur in a context of repetitive DNA sequence extending ≥ 2 kb. The approximately 2500 members of this family appear to be closely homologous, since no distant relatives are observed in the genome, even under relaxed conditions. Sequences of the 2108 family also belong to the long repeat class. However, this family differs from the 2034 family in that there are in the genome, besides the 20 closely related members referred to in Table 1, many additional sequences that are only distantly related to the CSp2108 probe sequence. These will be discussed in detail in a later paper. The large 2109 family consists mainly of short repeat sequence elements, and its members display degrees of relatedness ranging down to the lowest level permitted by the reaction conditions. This is indicated by the low average thermal stability of heteroduplexes between CSp2109 probes and genomic DNA (Table 1). Family 2109 has been separated into subfractions, a set of sequences reacting with a 180-nucleotide probe representing one end of the CSp2109 insert (2109A), and a set of sequences reacting with a probe containing the remaining 110 nucleotides (2109B). The genomic repeat sequence included in CSp2109 originally contained both the "A" and "B" portions (Posakony *et al.*, unpublished observations).

The three repeat families differ greatly in their representation in sea urchin RNAs (Table 1). Family 2034 is expressed mainly in intestine nuclear RNA; family 2108 transcripts occur mainly in oocyte RNA; and that portion of the 2109 sequence represented by the A probe is expressed primarily in intestine nuclear RNA, while transcripts homologous to the B probe are most prominent in gastrula nuclear RNA. The latter observations suggest that 2109 sequences homologous to the A portion of the repeat may occur in (transcribed) regions lacking the B element, and *vice versa*.

(b) *Frequency of occurrence of λ–sea urchin DNA recombinants bearing genomic members of the repetitive sequence families*

The probe repeat sequences described in Table 1 were labeled at the 5′ termini by the kinase procedure, and were used to screen the recombinant λ–sea urchin DNA genome libraries. The number of positive plaques obtained in these screens provides initial evidence on the distribution of the repeat family members in the genome. Thus, if the sequences belonging to a given family were widely distributed about the genome, the positive λ-recombinants would usually each contain only a single sequence element homologous to the repeat probe. The number of positive plaques expected is directly calculated for this case from the number of plaques screened and the reiteration frequency of the probe repeat family determined by renaturation kinetics. On the other hand, if in the genome the sequences of a repeat family occur in clusters, positive λ recombinants should often contain several copies, since the 10 to 18 kb length of the inserts is large compared to the length of most repeat sequences. The number of positive plaques expected would be correspondingly smaller.

Autoradiographs of 2109A and 2108 plaque screens are shown in Figure 1(a) and (b). Both plates contained about 2×10^4 p.f.u. While 343 plaques scored as positive with the 2109A tracer, only eight plaques reacted with the 2108 tracer. If all or most

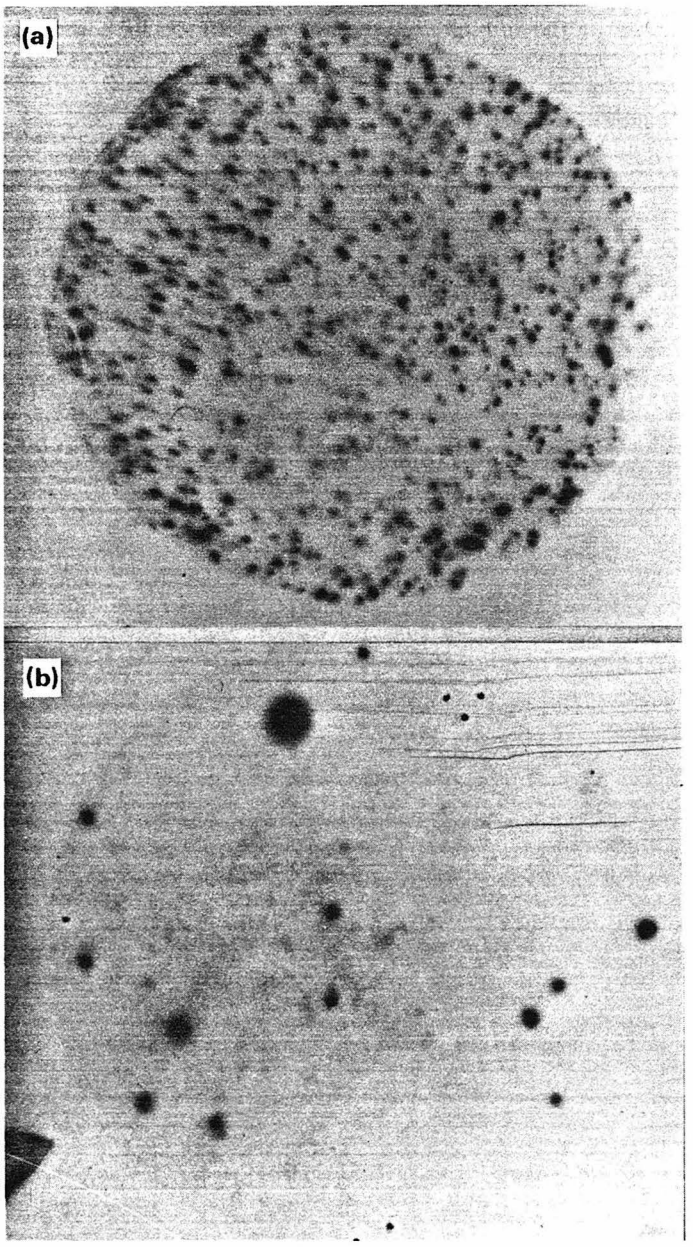


FIG. 1. Recombinant λ–sea urchin genome libraries screened with cloned repeat tracers. (a) Autoradiograph of a plate containing 2×10^4 p.f.u. from library SpλR_1A, and screened with the 2109A repeat probe described in Table 1. Hybridization was in 0·6 M-Na^+ at 55°C, and the screen was carried out by the procedure of Benton & Davis (1977), as described in Materials and Methods. (b) Autoradiograph of a plate containing 2×10^4 p.f.u. from library SpλH_3C, and screened with the 2108 repeat probe. Procedures were as for (a), except that hybridization and subsequent washing of the filter were carried out in 0·1 M-Na^+ at 70°C. This criterion suppresses reaction of distantly related members of the 2108 family.

members of these families occurred singly in the λ recombinants, the reiteration frequencies listed in Table 1 predict that about 330 clones per plate should have reacted with the 2109A probe, and about seven clones per plate should have reacted with the 2108 probe. Since these expectations are close to the observed values, this suggests that the majority of the members of each of these two repetitive sequence families are scattered in diverse regions of the genome, rather than clustered.

Data for screens with all four probe sequences are given in Table 2. Here it can be seen that screens carried out with the 2109B probe yield a result consistent with the 2109A experiment shown in Figure 1(a), and thus also indicate a non-clustered organization for the 2109 family. In contrast, the number of positive 2034 plaques is only about one-third that expected if members of this family were generally to occur singly. This implies an average of three of the 2034 sequences per positive λ-recombinant. Additional evidence suggesting multiple local clusters of 2034 family members in the λ recombinants is that many plaques react with the labeled probe sequence much more intensely than others. Since, as Table 1 shows, the members of this family are all closely related, this is likely to be due to differences in the number of copies per λ insert, rather than to variability in the degree of homology of the sequences.

When library screens were carried out with the 2108 probe at a more permissive criterion than used for the experiment shown in Figure 1(b), many more reactive plaques could be detected (not shown). These recombinants contained distant genomic relatives of the CSp2108 probe sequence. A random set of 2108 plaques was selected for rescreening and clonal purification which included examples of both

TABLE 2

Recovery of positive plaques in λ genome library screens with cloned repeat sequence probes

Repeat family	Positive plaques/ plate	Approximate expectation if repeat sequences occur singly[a]	Total no. λ recombinants plaque-purified[b]
2034	308	920	7
2108	8	7	15
2109 A probe	343	330	13
2109 B probe	357	370	27

Probes were obtained from the repeat plasmids CSp2108, CSp2109, and CSp2034 as described in the text. The length of the probe sequences is shown in Table 1. All plates contained 2×10^4 p.f.u./15 cm plate. Average data for 4 experiments are shown. Except for the 2108 experiment, which is described in the legend to Fig. 1, all screens for which data are presented here were carried out under the low criterion conditions specified in Materials and Methods.

[a] The expected number of positive plaques N is calculated for the assumption that the DNA of each recombinant λ that reacts with the probe sequence contains only 1 copy of this sequence. $N = PLR/G$, where P is the number of p.f.u./plate; L is the mean length of the sea urchin DNA insert in the λ recombinants, for which we take $1{\cdot}5 \times 10^4$ nucleotides; R is the repetition frequency of the repeat family, from Table 1; and G is the haploid genome size for *S. purpuratus*, or $8{\cdot}1 \times 10^8$ nucleotides.

[b] This column lists the total number of positive clones ultimately selected from each family for further analysis.

closely and distantly related sequences. The number of λ recombinants isolated for more detailed analysis of the sequence organization of all three repeat families is listed in the last column of Table 2.

(c) *Occurrence of the specific repeat sequence in λ genome library recombinants*

The number of sequence elements per λ recombinant that reacts with the cloned repeat probes was estimated by the gel blot method (Southern, 1975). The DNA was digested with various restriction enzymes, and the fragments were separated by agarose gel electrophoresis and blotted onto nitrocellulose filters for hybridization with the respective repeat tracers. Results from several representative experiments are shown in Figure 2. It can be seen in Figure 2(a), for example, that only one out of many restriction fragments reacts with the CSp2109B probe in each of two λ2109 recombinants (lanes A and B), while in contrast, the two λ2034 recombinants shown (lanes C and D) apparently contain multiple copies of this repeat, or elements thereof. The two λ2108 digests (lanes E and F) differ in that one contains a single band that reacts with the CSp2108 probe, while the other displays three reactive subfragments. In Figure 2(b) are reproduced several subdigests and blot hybridizations that were carried out on individual restriction subfragments initially found to contain reactive sequence(s). The examples shown include five cases in which the homologous repeat element is confined to only one relatively short subfragment (lanes A, B, C, E and G) and two cases in which evidence shown below indicates that there were clearly multiple copies of the relevant sequence clustered within the original reactive subfragment (i.e. λ2034-4, lane F; and λ2109B-16, lane D).

Since the repeat sequence elements are at least several hundred nucleotides in length, restriction sites will not infrequently occur within them. Indeed, primary sequence data (not shown) indicate that sites for some of the enzymes used in these experiments occur in the cloned repeat probes themselves, and it is probable that some members of the repeat family will contain the same sites. In a gel blot experiment, this would result in two (or more) bands for a single repeat sequence element. On the other hand, multiple homologous sequence elements could be present in any single reactive fragment large enough to contain them, as in the cases illustrated in Figure 2(b). To decrease the probability of misinterpretation due to either of these causes, we subjected each of the λ recombinants analyzed to digestion with a number of different restriction enzymes individually. Table 3 displays the results of these experiments, and includes our best estimates for the number of relevant repeat family members per whole sea urchin DNA insert. As shown there, sequences of the 2109 repeat family hybridized with either A or B tracers generally occur only once in a given λ recombinant. However, there are several exceptions. λ2109B-8 clearly contains at least two copies of the repeat, as does λ2109B-16, among others. In these particular cases, the multiple copies are known not to be contiguous to each other from mapping data, some of which are presented below. Though only a minority of the λ2109 recombinants include multiple copies of the repeat sequence, we note that the frequency with which they do occur is higher than would be expected if the elements of a 1000 member family were distributed completely at random throughout the genome.

The 2034 clones examined, except one (λ2034-14), contain multiple copies of this

TABLE 3

Number of restriction enzyme fragments bearing specific repeat family members in selected λ recombinants

Family	λ isolate[a]	Enzyme[b] *Hae*III	*Hha*I	*Hpa*II	*Hin*fI	Estimated number of family sequences/λ recombinant[c]
2109	λ2109A-6	2000	2500	2000	600	1
	8	1000	3000	700		1
	11	1000	2000	300	2500	2[d]
	12	500	3000	2000		1
	21	2500	3200	1500		1
	22	3500	1500	3000		(2)
			1000	2000		
	23	3500	3200	3000		1
				2000		
	24	3500	3000	3000		1
				2000		
	25	1000	2000	1800		1
			1500			
	26	3500	3500	2500		1
	27	1500	1000	3000		1
	28	1800	500	1500		1
			300			
	29	2500	300	1500		1
	λ2109B-6	900	600	800	1200	(2)
				600	1000	
	8	2300	4500	900	2000	2
		800	200	400	300	
	9	600	1700	2000	800	1
	10	700	700	2000	1100	1
	11	2000	1500	1500	2500	1
	14	1200	3000	2000	450	1
	15	2500	750	3000	1100	2
		750	200	2000	250	
					200	
	16	3500	4700	1500	2000	2
		650	700	1200		
				500		
	17	2500	2000	2000	2000	1
	18	500	2000			1
			500			
	30		200	1500		1
	31	2000	1200	3500		1
	32		2000	1100		1
	38		2000	3500		1
	39		3000	3500		1
	40		1600	2500		1
	42	2000	1500	2500		1
	44	3500	2000	3500		1
	45	3000	300	2000		1
	46	2500	1000	2000		1
	48	1800	2500	2000		1
	49	2500		1800		1
	50	1500	3500	2500		1
	52	3000	3500	2500		1
			1000			
	53	2000	3500	2000		1
	55	3000	2000	2500		1
	56	2000	3500	2500		1

TABLE 3 *(continued)*

2034	λ2034-3	2000	3000			(2)
		1000	2000			
	4	2500	2000	2000		4
		2400	1000	1500		
			750			
			500			
			500			
	9	2000	2000	3000		3
		700	500	1500		
		500	300	1000		
	10	1500		4000		4
		1300		3000		
		1200		2000		
		1100		900		
		800				
	13	2000	1000	2500		2
		1500	800	1000		
	14	1500		1000	1000	1
	18	3000	1000	2500		2
		2300	500	2000		
2108	λ2108-21	3000	2500	2000	3000	1
	32	3000	2500			1
High	33	3000	2500			1
criterion	34	3000	2500			1
selection	35	3000	2500			1
	36	3000	2500			1
	37	3000	2500			1
	λ2108-8	2500	1200	600	500	2
		1000		400		
		900				
	11	2000		1700	800	1
					600	
	12	1000	2500	2500	2000	2
Low		700		1500	850	
criterion	15	2000	1200	4000	2500	2
selection			500	1100		
			200			
	16	3000	2500	2000	3000	2
		800	2000		1000	
		350				
	17	3000		900	3000	3
		2500		800	2500	
				350	1500	
	18		600	1000	2000	(2)
				700	600	
	20	400		1000	1000	1
					850	

[a] All 2034 λ recombinants, 2108 recombinants 1–30, 2109A recombinants 1–20, and 2109B recombinants 1–40 were isolated from genome library $Sp\lambda R_1B$. The other λ recombinants listed were isolated from library $Sp\lambda H_3C$.

[b] Indicated are the lengths of the restriction enzyme fragments that react with the respective probes in gel blot hybridizations such as those shown in Fig. 2.

[c] The presence of 2 or more positive bands in 2 or more digests was taken to indicate the existence of multiple copies of the repeated sequence. Where 2 bands occur in only 2 digests, the conclusion is less certain, as indicated by parentheses.

[d] In this case, the listed fragments all occur twice within the insert.

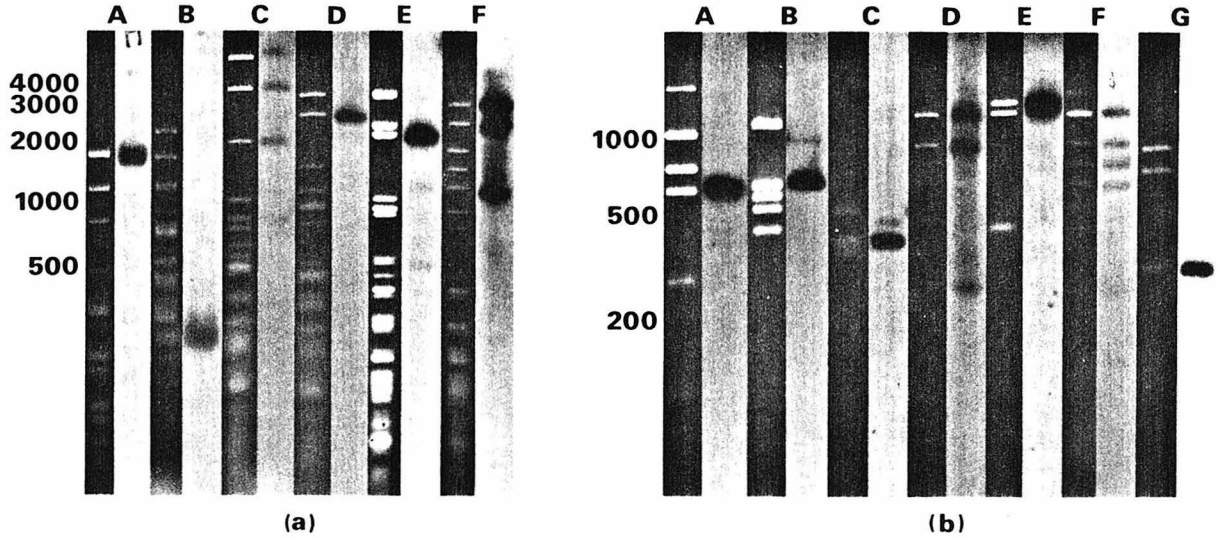


FIG. 2. Gel blot analyses of specific repeat sequences in selected λ recombinants. Isolates from the λ genome libraries were selected by screening with the cloned repeat probes as described in text. (a) The DNAs of the recombinant phage were digested with the indicated restriction enzymes and blotted for hybridization with the same probe fragment initially used to select the respective library isolates (see Materials and Methods for details). Each lane shows on the left the ethidium bromide fluorescence pattern of the digest, and on the right the autoradiograph of the corresponding gel blot. Fragment lengths are indicated along the ordinate. Lanes A and B: family 2109 isolates, digests reacted with [5′-^{32}P]2109B probe from plasmid CSp2109. A, λ2109B-9 DNA digested with *Hha*I; B, λ-2109B-14 DNA digested with *Hin*fI. Lanes C and D: family 2034 isolates, digests reacted with [5′-^{32}P]CSp2034 probe. C, λ2034-10 DNA digested with *Hpa*II; D, λ2034-13 DNA digested with *Hpa*II. Lanes E and F: family 2108 isolates, digests reacted with [5′-^{32}P]CSp2108 probe. E, λ2108-15 DNA digested with *Hpa*II; F, λ2108-17 DNA digested with *Hpa*II. (b) Subdigests and gel blots of single restriction enzyme fragments that reacted with the repeat probes in experiments such as that shown in (a). The appropriate fragments were eluted from agarose gels, redigested with the indicated restriction enzymes, and the digest again displayed on a gel. Restriction maps of many of the subfragments are shown in Fig. 3. As in (a), the left-hand track in each lane represents the ethidium bromide fluorescence pattern of the subdigest, while the right-hand track is the autoradiograph of the hybridized subdigest gel blot. Lanes A to E: subfragments from λ2109 isolates, reacted with [5′-^{32}P]2109A probe (A to C) or with [5′-^{32}P]2109B probe (D and E). A, *Hha*I subfragment of λ2109A-6, digested with *Hin*fI + *Hpa*II; B, *Hha*I subfragment of λ2109A-8, digested with *Hpa*II + *Hae*III. C, *Hha*I subfragment of λ2109A-11 digested with *Hpa*II + *Hae*III; D, *Hha*I subfragment of λ2109B-16, digested with *Hpa*II:E, *Hha*I subfragment of λ2109B-18, digested with *Hae*III. F, *Hha*I subfragment of λ2034-4, digested with *Hpa*II and hybridized with [5′-^{32}P]2034 probe. G. *Hha*I subfragment of λ2108-16, digested with *Hpa*II, and hybridized with [5′-^{32}P]2108 probe.

sequence. In the sample included in Table 3, the average number of these sequences per insert is 2·6, which can be compared to the estimate of three copies per insert derived from the data in Table 2. The agreement indicates that the presence of multiple copies of this repeat in the various 2034 λ recombinants does not induce frequent deletions, at least through the several rounds of replication required for each liquid culture amplification.

The 2108 family is a complex assemblage of variably related sequences, as noted above. With one exception, those λ recombinants selected at high criterion each contain only a single copy of the 2108 sequence. Though the particular restriction fragments that include this sequence are all the same in size (Table 3), the inserts of the various high criterion λ2108 recombinants are diverse with respect to their overall restriction digest pattern. It follows that the high criterion 2108 sequence element is part of a longer repeat unit that occurs in many different regions of the genome, in different sequence environments. This inference is supported in detail by Scheller *et al.* (unpublished results). The λ2108 recombinants selected at a more permissive criterion occur in one, two or three copies per insert. The degree of relatedness and the actual size of this "superfamily" of diverse sequences is not well known, and the statistical significance of the multiple local occurrences of these sequences remains uncertain.

(d) *Sequence environment around repeat family members*

The genomic repetition frequency of sequences flanking several members of each repeat family was measured. While most repeats in the sea urchin genome are short and are surrounded by single copy regions, data from other sources (summarized in Table 1) indicated that the 2108 and 2034 families belong to the long repeat class. Thus, it is expected that, on the average, the sequences immediately flanking the elements reacting with these two probes in the λ recombinants would be repetitive as well. There was no previous information regarding either the repeat length or the sequence environment characteristic of the 2109 family.

Restriction subfragments bearing members of the three repeat families were isolated from a number of the λ recombinants. Figure 3 displays the location of various enzyme sites within these subfragments, and indicates the positions of the specific repeat sequences (▼). In most cases there is only one such sequence element within the several thousand nucleotides of the mapped fragment, as expected from the data of Table 3, for at least the majority of 2109 and 2108 examples. However, the restriction maps provide interesting additional information regarding the arrangement of the 2034 repeat sequences. At least two different subfragments bearing these repeats were isolated and mapped from λ2034-4, λ2034-9, and λ2034-10. Figure 3 shows that all of the mapped subfragments from these recombinants contain multiple copies of the 2034 repeat. In the case of λ2034-4, the two mapped subfragments are identical in the pattern of restriction sites except for one *Hae*III site. All the other 2034 restriction maps are unique. These data suggest that the genome includes both tandem 2034 sequences, some of which are almost exact replicas, and tightly clustered but non-exact replicas of regions carrying 2034 sequences, as well as the somewhat more widely spaced 2034 sequence clusters indicated in Table 3. Figure 3 also reveals a 2109 sequence cluster (λ2109B-16).

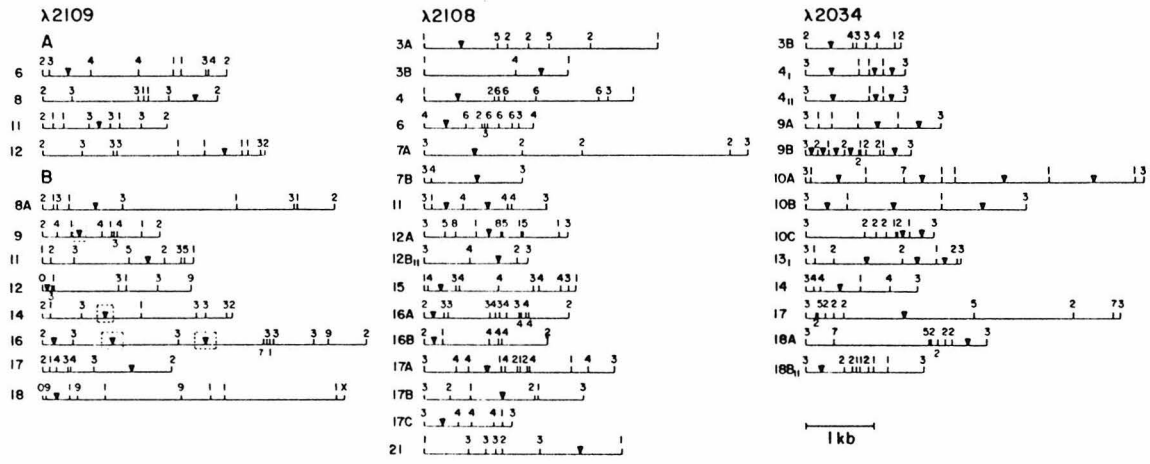


FIG. 3. Restriction maps of subfragments from λ recombinants bearing repeat family members. Subfragments were obtained from the various λ recombinant DNAs after digestion with the restriction enzymes indicated at the termini. The subfragments shown were mapped by partial and double digests with other enzymes: 1, *Hae*III; 2, *Hha*I; 3, *Hpa*II; 4, *Hin*fI; 5, *Hin*cII; 6, *Taq*I; 7, *Ava*I; 8, *Ava*II; 9, *Hin*dIII; 0, *Eco*RI; X, *Xba*I. Where 2 subfragments bearing members of the repeat family in a given λ recombinant are released by the same enzyme, these are denoted A and B in order of decreasing size. The position of the relevant repeat family member(s) (▼) was determined by gel blot hybridization as described in the previous section of this paper. The location of the repeat sequence within the region defined by the nearest restriction sites is unknown. Locations of 2 of the 2109B sequence elements observed in λ2109B-16, 14, and 9 are specified by the dimensions of the heteroduplex shown in Fig. 5(d), and their locations are noted here by the boxed regions.

These elements are neighboring but not contiguous, as shown later in this paper. This suggests that a repeat family in which most members occur singly throughout the genome may also include local clusters of several sequence elements.

Genomic reiteration frequencies were determined within the mapped regions by isolating and labeling the appropriate DNA subfragments with ^{32}P, and reacting them with excess sea urchin DNA. Single copy [^{3}H]DNA kinetic standards were included in each reaction. Figure 4 shows representative examples of renaturation kinetics for the relevant regions of three λ recombinants. Reactions carried out with a set of subfragments from λ2109B-9 are illustrated in Figure 4(a). The 2109 family member on subfragment A reacts 680 times faster than the internal single copy standard, which is consistent with the overall repetition frequency estimated previously for this family (Table 1). All of the flanking fragments clearly react as

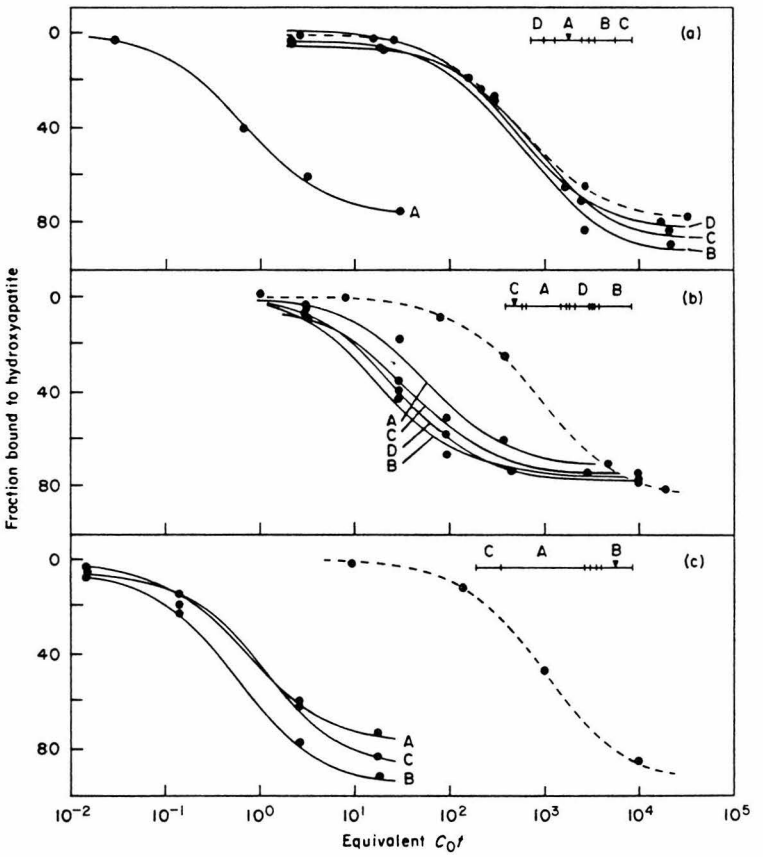
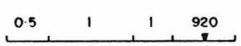
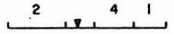
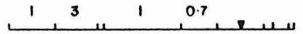
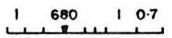
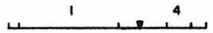
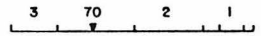
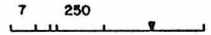
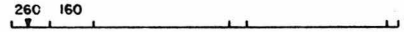


FIG. 4. Renaturation of subfragments from λ recombinants with excess sea urchin DNA. The subfragments indicated in the restriction map in each panel (see Fig. 3 for enzyme sites) were labeled with ^{32}P by the kinase reaction, and reacted with sheared sea urchin DNA in the presence of a [^{3}H]DNA single copy tracer (broken lines). Details are given in Materials and Methods. The subfragment indicated by ▼ includes the relevant repeat sequence family member. Individual kinetic measurements are indicated by the letters that denote the restriction fragments tested. The rate constants for these reactions are shown in Table 4. (a) *Hha*I subfragment from λ2109B-9; (b) *Hha*I A subfragment from λ2108-16; (c) *Hpa*II A subfragment from λ2034-18.

TABLE 4

Reiteration frequencies for regions surrounding specific repeat sequence elements

λ recombinant	Digest[a]	Fragment length	K[b]	K corrected[c]	Genomic reiteration frequency[d]	Restriction map[e]
λ2034-9	*Hae*III	640	4·4	4·1	3280	
		640	4·4	4·1	3280	3200 3280 3280
		560	4·0	4·0	3200	
λ2034-17	*Hha*I + *Hin*fI	1400	3·7	1·6	1260	
		1300	3·5	1·6	1260	1260 1260 950
		750	1·5	1·2	950	
λ2034-18	*Hha*I	1380	1·5	0·65	780	
		500	1·6	1·8	2160	1360 780 2160
		380	0·90	1·13	1360	
λ2108-11	*Hin*fI	640	0·015	0·014	13	
		560	0·02	0·021	19	19 13 12
		480	0·012	0·013	12	
λ2108-12	*Hae*III	720	0·004	0·0033	4	
		680	0·004	0·0035	4	4 4 4
		520	0·003	0·0032	4	
λ2108-15	*Hin*fI	570	0·008	0·009	9	
		510	0·02	0·021	21	21 9 6 1
		400	0·005	0·006	6	
		315	0·001	0·0013	1	
λ2108-16	*Hin*fI + *Hpa*II	600	0·02	0·02	20	
		580	0·052	0·053	53	40 20 70 53
		240	0·025	0·04	40	
		160	0·036	0·07	70	
λ2108-21	*Hpa*II	1170	0·10	0·051	51	
		780	0·047	0·036	36	30 36 51
		550	0·028	0·03	30	
λ2109A-6	*Hpa*II + *Hin*fI	995	0·0027	0·0016	1	
		680	0·0064	0·0056	4	410 4 1 1
		600	0·53	0·53	410	
		260	0·001	0·0015	1	

λ2109A-8	*Hpa*II	950	0·0025	0·0016	1	
		700	1·4	1·2	920	
		440	0·0011	0·0012	1	
		410	0·0006	0·0007	0·5	
λ2109A-11	*Hpa*II	670	0·0024	0·0021	2	
		450	0·0045	0·0052	4	
		360	0·0012	0·0015	1	
λ2109A-12	*Hae*III + *Hpa*II	900	0·0011	0·00073	1	
		560	0·0011	0·0011	1	
		440	0·0033	0·0039	3	
		400	0·00071	0·0009	0·7	
λ2109B-9	*Hae*III + *Hin*fI	440	1·4	1·63	680	
		360	0·0016	0·0021	1	
		340	0·0013	0·0017	0·7	
		210	0·002	0·0033	1	
λ2109B-11	*Hha*I + *Hin*cII	1120	0·0038	0·0020	1	
		305	0·0044	0·0061	4	
λ2109B-14	*Hpa*II + *Hae*III	850	0·16	0·11	70	
		800	0·005	0·0038	2	
		300	0·0035	0·0049	3	
		150	0·001	0·002	1	
λ2109B-17[c]	*Hpa*II + *Hin*fI	1070	—	—	—	
		380	—	—	250	
		300	—	—	7	
λ2109B-18	*Hae*III	500	0·158	0·173	160	
		440	0·24	0·280	260	

[a] The restriction enzyme used to digest the subfragment whose map is shown in the last column. For other enzyme sites and further identification of the subfragments, see Fig. 3.

[b] K is the experimentally determined 2nd-order rate constant and is given in units of $M^{-1}\ s^{-1}$.

[c] The corrected rate is the rate of reassociation after correcting for the disparity in length between the driver DNA (600 nucleotides) and the tracer fragment (Chamberlin *et al.*, 1978).

[d] The reiteration frequency is obtained by dividing the corrected fragment rate by the rate of reassociation of the internal single copy standard, which varied slightly from experiment to experiment, and averaged about 0·0014 $M^{-1}\ s^{-1}$. The accuracy of these estimates is limited by: (1) the small number of data points necessitated by the large number of samples run; (2) the effects of lack of complete homology where repeat sequences are reacted. This error cannot be specifically evaluated for the flanking repeats, since in most cases, thermal stability estimates for the reaction products were not determined. Despite the presence of the internal single copy standard, these factors could induce a 2 to 3-fold error in the estimates given, except for single copy sequences, and except for 2034 sequences, which are in general highly homologous.

[e] Reiteration frequencies measured by the titration method (Scheller *et al.*, 1978; and unpublished results).

single copy sequences. In contrast, the measurements shown in Figure 4(b) and (c) demonstrate the repetitive nature of the sequence environments of a 2108 repeat in λ2108-16, and of a 2034 family member in λ2034-18.

Results of a number of experiments of this nature are summarized in Table 4. It is clear from these data that the 2109 family consists primarily of short (i.e. several hundred nucleotides) repeat sequences interspersed in regions of single copy DNA. Two individual exceptions are found in λ2109B-17 and λ2109B-18, in which both of the fragments tested were moderately repetitive. In several cases, the rate of reaction of the flanking sequences was slightly faster than that of the single copy standard, suggesting that one or two related sequences exist somewhere within the sea urchin genome. Table 4 shows that the repeat elements of both the 2108 and 2034 families are usually, though not invariably, flanked by sequences reiterated in the genome to about the same extent.

These determinations of reiteration frequency are probably subject to at least a twofold error (see the legend to Table 4). The significance of even some of the larger rate differences observed within the same region (e.g. in λ2108-16) is therefore not clear. If real, these differences could indicate that some of the contiguous repeat sequence elements occur separately, elsewhere in the genome. In one case, the mapped subfragment appears to include a terminal junction of the 2108 long repeat. Thus, all the tested sequences on the left end of the λ2108-15 subfragment are repetitive, while in the orientation shown, the right end is single copy.

(e) *Heteroduplex analysis of λ2109 sequences*

In Figure 5 are shown electron micrographs of heteroduplexes formed between DNA molecules from different λ2109 recombinants. The structures observed are consistent with the conclusions drawn above regarding the organization of this family. In Figure 5(a), the complete inserts of λ2109A-22 and λ2109A-24 can be observed extending between the forks that represent the beginning of the right and left arms of the Charon 4 vector. The single duplex structure within the sea urchin DNA must be the 2109 sequence, though there are almost certainly many other repeat sequences in both of these long inserts. No other complementary region can be seen. The duplex structure is 210 ± 25 nucleotides long (Fig. 5(b)). Figure 5(c) displays a heteroduplex between restriction subfragments from λ2109B-9 and λ2109B-14. The only repetitive sequence in both these subfragments is located where the 2109B probe reacts, since the remainders of both subfragments are single copy or nearly single copy (Table 4). Thus, the identification of the heteroduplex as a 2109 repeat is in this case unequivocal. The length of the homologous region between the four non-homologous single copy tails is again about 200 nucleotides. Figure 5(d) shows the *Hha*I subfragment of λ2109B-16, reacting with itself at an homologous but out of register site. This is almost certainly the 2109 sequence, since the map of this fragment in Figure 3 indicates that it contains at least two 2109 repeat sequence elements. Figure 5(e) shows that the same length of duplex occurs in about the same position relative to the ends of the λ2109B-16 subfragment when this subfragment is reacted with the λ2109B-14 subfragment used in Figure 5(c). The experiment thus confirms the unusual double occurrence of the 2109 repeat in the local region of the genome included in λ2109B-16. This fragment therefore

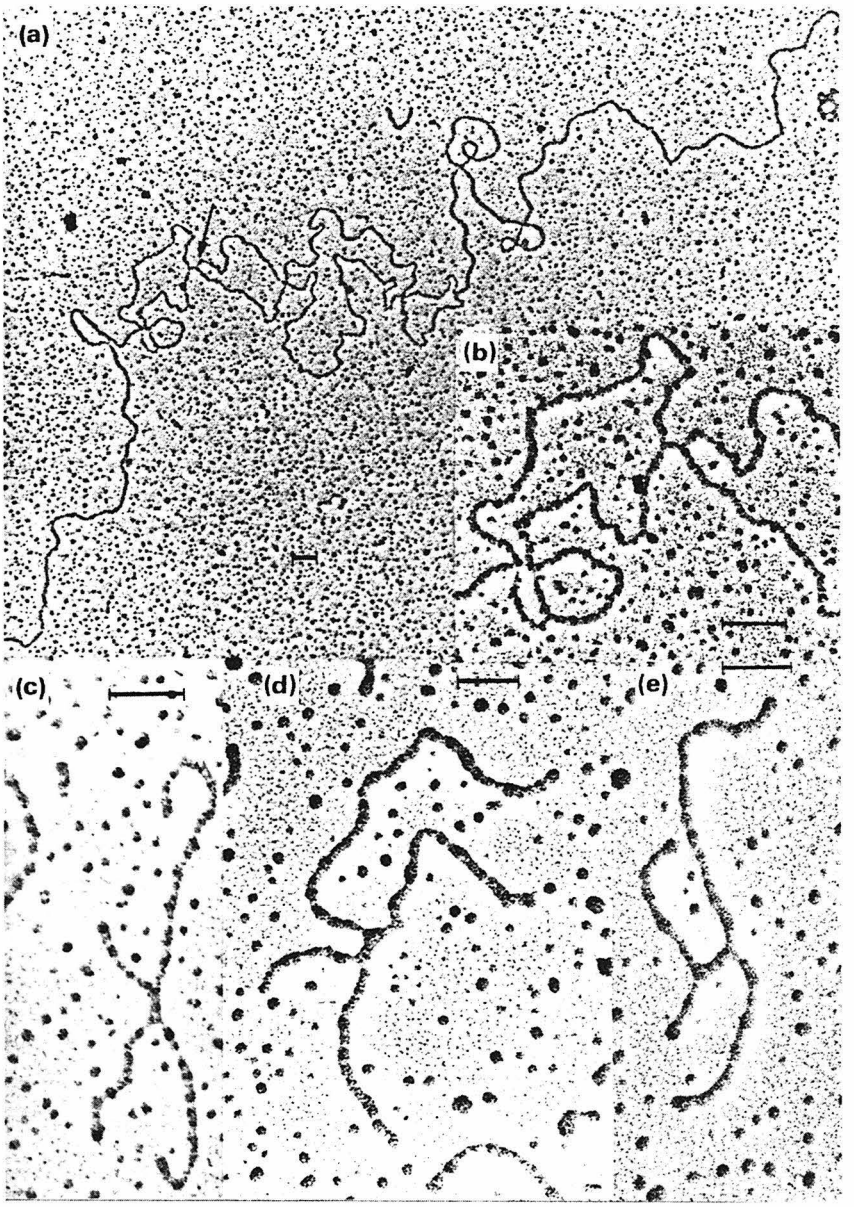


FIG. 5. Electron micrographs of heteroduplexes formed between homologous regions of 2109 λ recombinants. (a) Heteroduplex between complete genomes of λ2109A-22 and λ2109A-24. The arrow indicates the homologous region between the sea urchin DNA inserts. The phage were mixed, treated with alkali and, after neutralization, hybridized as described by Davis *et al.* (1971). The DNA was spread from a hypophase containing 55% formamide and hyperphase containing 25% formamide (Davis *et al.*, 1971). (b) Further magnification of the heteroduplex in (a). (c) Heteroduplex between 1·7 kb *Hha*I subfragment of λ2109B-9 and 2·85 kb *Hha*I subfragment of λ2109B-14. Restriction maps and reiteration frequencies for these subfragments may be found in Fig. 3 and Table 4. (d) Out of register duplex formed in self-reaction of 4·7 kb *Hha*I subfragment from λ2109B-16. This subfragment contains at least two 2109B sequence elements (Fig. 3). (e) Heteroduplex between 4·7 kb *Hha*I subfragment of λ2109B-16 and 2·85 kb *Hha*I subfragment of λ2109B-14. The length of the bar in (b) to (e) represents 0·5 kb.

contains two copies of the 2109B repeat that are separated by 1000 nucleotides of other DNA sequences.

4. Discussion

(a) *Organization of the three repeat families*

The 2109 repeat family includes about 1000 reasonably well-matched sequences, though additional, more distant relatives exist in the genome. The repetitive sequences of this family are organized for the most part as 200 to 300-nucleotide elements interspersed in single copy DNA. Most individual family members occur distantly from each other, since few of the 10 to 18 kb inserts in the 2109 λ recombinants include more than a single 2109 sequence, and since the number of recombinants reacting with the 2109 probe in genome screens correctly predicts the size of this family. However, there is a minor fraction of 2109 family members that are located only a short distance from other 2109 repeats. The fraction of 2109 family members in the genome that occurs in multiple local arrays may be overestimated in the set of 2109 λ recombinants, since their stronger screening signals could have resulted in a bias toward their selection. We cannot exclude the alternative that there is actually a more clustered 2109 sequence organization in the genome than we observed if such arrangements are unstable in the λ recombinants. However, we regard this as highly unlikely because of the quantitative agreement between the reiteration frequency of the 2109 family measured in whole DNA and the screening results shown in Table 2. Furthermore, many recombinants containing clustered copies of 2034 repeat sequences have been recovered and studied without difficulty.

The 2108 sequence is an element of a longer repeat that extends for several thousand nucleotides, as shown for several examples in Table 4. Some members of this family studied by Scheller *et al.* (unpublished results) are about 4·5 kb in length. This result is consistent with the observation (Moore *et al.*, 1980) that the 2108 probe reacts preferentially with a long repeat fraction of the genome. The 2108 family is organized in a complex and interesting way. When reactions are carried out under fairly stringent criterion conditions, only about 20 sequences closely related to the CSp2108 probe are observed. However, many more distant relatives, constituting a large "superfamily", exist in the genome. Differences in the apparent reiteration frequencies between several of the 2108 sequences in Table 4 are due to the fact that these sequences belong to different subsets of the 2108 superfamily. Despite this complication, it is clear from the experiments shown in Table 3 that most (or all) of the 20 members of the closely related 2108 sequence subfamily studied here occur singly. That is, they reside in regions of the genome that are at least 10 to 20 kb apart. Table 3 shows that the large number of less closely related members of the 2108 superfamily must also occur in many different locations.

The 2034 sequence is also an element of a long repeat, but members of this family are all closely related. The amount of sequence divergence between various 2034 family members (about 4%) is no greater than between average single copy sequences of different individual *S. purpuratus* genomes (Britten *et al.*, 1978). This large family has about 2500 members and these sequences are partially clustered

and partially scattered about the genome. At one extreme are regions such as those in Figure 3, which include tandem or tightly clustered 2034 sequences. However, most clusters cannot contain many copies of the 2034 sequence, since the number of reactive restriction subfragments per λ recombinant is not very large (Table 3), and since the 2034 genome screens revealed as many as one-third of the number of positive plaques expected on the basis of single occurrences per insert (Table 2). Thus, there are probably several hundred small clusters of 2034 sequences in the genome, as well as some single occurrences (e.g. in λ2034-14). In addition, the genome contains one or more very large deposits that are probably tandemly arranged. The evidence for this is presented elsewhere by Moore *et al.* (1980), who noticed a prominent band in 2034 genome blots that includes a significant though small fraction of the 2034 sequence. The restriction pattern and tandem sequence organization expected for an element derived from this genomic repository of 2034 sequences is displayed by the subfragment of λ2034-4 shown in Figure 3.

(b) *Global distribution of members of individual repeat families*

The major finding reported here is that the members of all three repeat families populate many separate regions of the genome. Suppose, for example, that the organization of the 1000-member 2109 family in the genome is reasonably represented in the set of λ recombinants listed in Table 3. There are 40 single occurrences of 2109 sequences per average 15 kb insert, plus seven double occurrences. It follows that even on the extreme assumption that all 2109 sequences are as close as allowed by these statistics, (i.e. about one 2109 sequence every 20 kb), this short repeat family would be dispersed over no less than 2×10^7 nucleotides of DNA. Thus, unless the λ recombinants are not representative, the 2109 family could not be confined to one or a few chromosomal "domains" of the size observed in *Drosophila* cells (Benyajati & Worcel, 1976), or of the size of polytene chromosome bands. Most likely, the sequences of the 2109 family are dispersed over a large portion of the genome, although this is not shown. It is also clear from the foregoing that long repeats, such as the 2108 and 2034 sequences, are widely distributed in the genome. *In situ* hybridization has demonstrated extensive genomic dispersion of several cloned *Drosophila* long repeats as well (Finnegan *et al.*, 1978; Potter *et al.*, 1979; Wensink *et al.*, 1979; Rubin *et al.*, 1980).

Interspersed short repeats and their flanking sequences account for a major portion of the genome in organisms such as *Xenopus* (Davidson *et al.*, 1973; Chamberlin *et al.*, 1975), human (Schmid & Deininger, 1975), and sea urchin (Graham *et al.*, 1974). An interesting and relevant example from a particular region of a mammalian genome is the rabbit β globin gene cluster. Shen & Maniatis (1980) showed that the 44 kb of DNA that includes the four β globin genes contains at least 20 repeat sequence elements interspersed among other sequences, and representing a minimum of four distinct families. Our present findings illuminate an important consequence of repetitive sequence interspersion. Each interspersed repeat family organized like the 2109 family endows the genome with a network of sequence relationships that links physically distant single copy regions of the DNA. In the sea urchin genome there are perhaps half a million short repeat sequence elements, belonging to at least 1000 non-homologous families. The matrix of relationships

constructed by these sequences extends throughout most of the genome, since only a minor fraction of the single copy DNA is devoid of them.

The concept of genomic sequence organization advanced here is abstract, and is based solely on structural information. It may or may not have relevance to the co-ordination of genome function. However, it is clearly important in considering the processes by which the genome has evolved.

(c) *The evolutionary process of repeat sequence dispersion*

In previous articles, we have drawn attention to the possibility that evolutionary mechanisms exist for the insertion (and disappearance) of interspersed repetitive sequences. This is directly implied by a recent study (Moore *et al.*, 1978), which showed that some repeat sequence families differ markedly in size among related species of sea urchins. One way of envisioning such a process is as follows. At an early stage in the growth of a repeat sequence family, one or a few potentially large blocks of tandem repeats are copied from a pre-existing sequence. A mechanism then begins to function by which the length of these blocks is progressively decreased, as fragments of them are excised and inserted elsewhere in the genome. For example, out of register pairing might occur, followed by excision of non-aligned sequences, some of which might then reinsert at random non-homologous locations by the same kinds of reactions as are responsible for insertion of foreign DNA elements in DNA transformation experiments (Wigler *et al.*, 1979). This process would ultimately result in a family of singly occurring repeats interspersed in many regions of the genome. Eventually, decay of recognition among the separated family members would be likely to occur, by means of sequence divergence, deletion or internal reorganization.

Viewed in light of these speculations, the 2034 family seems to be in an early stage of expansion and dispersion in the genome of *S. purpuratus*. The large repository of 2034 sequences cited above (Moore *et al.*, 1980) is evidently of recent origin, since this set of a very similar and probably tandem repeats is absent from the genome of *Strongylocentrotus franciscanus*, even though the repeat family itself is well-presented in the latter species. Most other 2034 sequences now in the *S. purpuratus* genome are scattered about in clusters, and only a few are as yet singly interspersed. These clustered sequences are probably also of recent origin, given the low overall degree of sequence divergence in this family (Klein *et al.*, 1978). In contrast, the 2109 family appears to be at an advanced stage of its evolutionary dispersion, though a few nearly contiguous 2109 sequences still exist in the genome, since one example was encountered in our scan of 2109 family recombinants. The relatively large differences between the sequences of many 2109 family members (Klein *et al.*, 1978; Posakony *et al.*, unpublished results) suggest that the replication event(s) giving rise to the present families occurred much longer ago for the 2109 family than for the 2034 family.

Repeat family size and organization are remarkably plastic. This raises the question of how evidence for specificity of repeat family expression (e.g. see Federoff *et al.*, 1977; Costantini *et al.*, 1978; Scheller *et al.*, 1978; Moore *et al.*, 1980) is to be interpreted. A possible answer is that some repeat family members are transcribed and are in some sense functional, while particularly in cases of large, rapidly

expanding families, many of the "newer" copies are not. The networks of homology within the genome resulting from evolutionary repeat family dispersion may thus include both presently expressed sets of repeat sequences, and sets that may potentially be deleted or in some cases included in useful patterns of expression in future evolutionary time.

We thank Dr Elliot Meyerowitz for his helpful and critical review of this manuscript. This research was supported by National Institutes of Health grant GM-20927. One author (D. M. A.) was supported by a National Institutes of Health postdoctoral fellowship (HD-05510), two authors (R. H. S. and J. W. P.) were supported by a National Institutes of Health National Research Service award (GM-07616), and one author (L. B. M.) was supported by a California Institute of Technology summer undergraduate research fellowship.

REFERENCES

Benton, W. D. & Davis, R. W. (1977). *Science*, **196**, 180–182.

Benyajati, C. & Worcel, A. (1976). *Cell*, **9**, 393–407.

Blattner, F. R., Williams, B. G., Blechl, A. E., Denniston-Thompson, K., Faber, H. E., Furlong, L.-A., Grunwald, D. J., Kiefer, D. O., Moore, D. D., Schumm, J. W., Sheldon, E. L. & Smithies, O. (1977). *Science*, **196**, 161–169.

Britten, R. J., Graham, D. E. & Neufeld, B. R. (1974). In *Methods in Enzymology* (Grossman, L. & Moldave, K., eds), pp. 363–406, Academic Press, New York.

Britten, R. J., Graham, D. E., Eden, F. C., Painchaud, D. M. & Davidson, E. H. (1976). *J. Mol. Evol.* **9**, 1–23.

Britten, R. J., Cetta, A. & Davidson, E. H. (1978). *Cell*, **15**, 1175–1186.

Chamberlin, M. E., Britten, R. J. & Davidson, E. H. (1975). *J. Mol. Biol.* **96**, 317–333.

Chamberlin, M. E., Galau, G. A., Britten, R. J. & Davidson, E. H. (1978). *Nucl. Acids Res.* **5**, 2073–2094.

Costantini, F. D., Scheller, R. H., Britten, R. J. & Davidson, E. H. (1978). *Cell*, **15**, 173–187.

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

Davidson, E. H., Galau, G. A., Angerer, R. C. & Britten, R. J. (1975). *Chromosoma*, **51**, 253–259.

Davis, R. W., Simon, M. & Davidson, N. (1971). In *Methods in Enzymology* (Grossman, L. & Moldave, K., eds), pp. 413–428, Academic Press, New York.

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

Eden, F. C., Graham, D. E., Davidson, E. H. & Britten, R. J. (1977). *Nucl. Acids Res.* **4**, 1553–1567.

Federoff, N., Wellauer, P. K. & Wall, R. (1977). *Cell*, **10**, 597–610.

Finnegan, D. J., Rubin, G. M., Young, M. W. & Hogness, D. S. (1978). *Cold Spring Harbor Symp. Quant. Biol.* **42**, 1053–1063.

Galau, G. A., Klein, W. H., Davis, M. M., Wold, B. J., Britten, R. J. & Davidson, E. H. (1976). *Cell*, **7**, 487–505.

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

Klein, W. H., Thomas, T. L., Lai, C., Scheller, R. H., Britten, R. J. & Davidson, E. H. (1978). *Cell*, **14**, 889–900.

Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K. & Efstratiadis, A. (1978). *Cell*, **15**, 687–701.

Maxam, A. M. & Gilbert, W. (1977). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 560–564.

Moore, G. P., Scheller, R. H., Davidson, E. H. & Britten, R. J. (1978). *Cell*, **15**, 649–660.

Moore, G. P., Costantini, F. D., Posakony, J. W., Britten, R. J. & Davidson, E. H. (1980). *Science*, **208**, 1046–1048.

Potter, S. S., Brorein, W. J., Dunsmuir, P. & Rubin, G. M. (1979). *Cell*, **17**, 415–427.

Rubin, G. M., Brorein, W. J., Dunsmuir, P., Flavell, A. J., Levis, R., Strobel, E., Toole, J. J. & Young, E. (1980). *Cold Spring Harbor. Symp. Quant. Biol.* **45**, in the press.

Scheller, R. H., Thomas, T. L., Lee, A. S., Klein, W. H., Niles, W. D., Britten, R. J. & Davidson, E. H. (1977). *Science*, **196**, 197–200.

Scheller, R. H., Costantini, F. D., Kozlowski, M. R., Britten, R. J. & Davidson, E. H. (1978). *Cell*, **15**, 189–203.

Schmid, C. W. & Deininger, P. L. (1975). *Cell*, **6**, 345–357.

Shen, C.-K. J. & Maniatis, T. (1980). *Cell*, **19**, 379–391.

So, M., Gill, R. & Falkow, S. (1975). *Mol. Gen. Genet.* **142**, 239–249.

Southern, E. M. (1975). *J. Mol. Biol.* **98**, 503–517.

Wensink, P. C., Tabata, S. & Pachl, C. (1979). *Cell*, **18**, 1231–1246.

Wigler, M., Sweet, R., Sim, G. K., Wold, B. J., Pellicer, A., Lacy, E., Maniatis, T., Silverstein, S. & Axel, R. (1979). *Cell*, **16**, 777–785.

Yamamoto, K. R., Alberts, B. M., Benzinger, R., Lawhorne, L. & Treiber, C. (1970). *Virology*. **40**, 734–744.

Chapter II

Reprinted from *J. Mol. Biol.* (1981) **149**, 41–67

Repetitive Sequences of the Sea Urchin Genome

III. Nucleotide Sequences of Cloned Repeat Elements

J. W. Posakony, R. H. Scheller, D. M. Anderson, R. J. Britten and E. H. Davidson

J. Mol. Biol. (1981) **149**, 41–67

Repetitive Sequences of the Sea Urchin Genome

III. Nucleotide Sequences of Cloned Repeat Elements

J. W. Posakony, R. H. Scheller, D. M. Anderson, R. J. Britten and E. H. Davidson

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

(Received 5 January 1981)

The nucleotide sequences of eight randomly selected, cloned, repetitive sequence elements were determined. No homologies exist among the eight sequences that are sufficient to promote cross-reaction between them under standard conditions of measurement. Thus, each sequence is representative of a different repeat sequence family. Statistically significant but short (8 to 41 nucleotides) internal direct and inverse repetitions occupy a minor fraction of the sequence length in five of the eight repeat sequences. None contains internal reverse repeats sufficiently long to permit inclusion in the "foldback" DNA fraction. The general lack of internal sequence homology means that the sequence complexity of the eight clones is approximately equal to the length of the cloned inserts. Nucleotide sequences of three different members of one particular short interspersed repeat sequence family are also reported. Comparison of these sequences reveals that both the number and order of internal sequence subelements differ among family members. The results show that both fine-scale rearrangement and sequence divergence have occurred during the evolution of this repeat family.

1. Introduction

Moderately repetitive sequences are dispersed throughout the genomes of many animals and constitute a significant fraction of the DNA in most species. Many of these sequences are only a few hundred nucleotides long and are terminated by other DNA sequences, which permits their easy identification and isolation. Yet little information regarding the character of the nucleotide sequences in moderately repetitive DNA has appeared. In this paper we report sequences of eight cloned repeats, randomly selected from a large set of plasmid recombinants. Each of these contains a replica of a single renatured repetitive DNA duplex from the sea urchin genome. The basic characteristics of these repeat clones, and of the genomic sequence families to which they belong, were reported earlier (Scheller *et al.*, 1977*a*: Klein *et al.*, 1978). The organization and characteristics of several of these repeat families have been reported in some detail in the previous papers of the present series (Anderson *et al.*, 1981; Scheller *et al.*, 1981). All of the cloned repeats whose

0022-2836/81/170041-27 $02.00/0

sequences are presented here are known to be represented in sea urchin egg poly(A) RNA, and are also transcribed to various extents into nuclear RNAs, in a stage and tissue-specific manner. We regard these eight repetitive sequences as fairly typical of sea urchin genomic repeats, though of course the variety present in this class of sequences is so great that such a small sample can only be regarded as indicative. Nonetheless, we can draw some simple though important conclusions. Most of the eight sea urchin repeats are not translatable sequences; a relatively small fraction of their length consists of internal repetition, either direct or inverse; and they are not related to each other. These results imply directly that sea urchin repetitive DNA consists of an assemblage of many distinct repeat sequence families, as deduced earlier from renaturation kinetics (Graham *et al.*, 1974; Klein *et al.*, 1978; Costantini *et al.*, 1980).

We have also compared the sequences of three cloned short repeat elements from different regions of the genome, all belonging to a specific interspersed repeat sequence family, the 2109 family (Anderson *et al.*, 1981). These repeats turn out not to be colinear and instead display alternative arrangements of internal sequence subelements. This surprising result focuses attention on relatively short sequence subelements, <40 nucleotides in length, in considering both the evolution and the function of interspersed repetitive sequences.

2. Materials and Methods

Detailed methods used in this laboratory for preparation of bacteriophage and plasmid recombinant DNAs, gel blots, electron microscope heteroduplex analysis and associated procedures were described by Anderson *et al.* (1981) and Scheller *et al.* (1981).

(a) *Preparation of 5′ terminally labeled DNA fragments for sequencing*

DNAs of plasmid clones containing individual sea urchin repetitive sequence elements (Scheller *et al.*, 1977*a*; Klein *et al.*, 1978) were cut with *Eco*RI to release the inserted fragment(s). DNA of clone CSp2034 was also treated with *Xba*I to yield 2 sea urchin DNA fragments which were sequenced separately. Restriction fragments 2 to 5 kb† long, previously identified as sites of sequences homologous to the CSp2109B repeat, were isolated from genomic λ recombinants and further digested to yield shorter fragments suitable for DNA sequencing. These fragments included all sequence elements reacting with the CSp2109B probe. Specifically, the 2·2 kb *Eco*RI-*Hin*dIII fragment from λ2109B-12 (see Anderson *et al.*, 1981) was digested with *Hin*fI to yield a 270 nucleotide fragment for sequencing, and the 4·4 kb *Eco*RI-*Xba*I fragment from λ2109B-18 (see Anderson *et al.*, 1981) was digested with *Xho*I to yield a 190 nucleotide fragment. After isopropanol precipitation, the digested DNAs were dephosphorylated at their 5′ termini with bacterial alkaline phosphatase and labeled by the T4 polynucleotide kinase reaction (Maxam & Gilbert, 1977; Anderson *et al.*, 1981). These reaction mixtures were either precipitated directly with isopropanol or chromatographed on Sephadex G-50 columns to remove remaining [γ-^{32}P]ATP. For strand separation the ^{32}P-labeled DNA was dissolved in 0·1 × Tris/borate/EDTA buffer (Maxam & Gilbert, 1977) plus 10% glycerol, heat denatured in boiling water for 3 min, and quenched in an ice/water bath. Electrophoresis was carried out on 4% to 6% polyacrylamide slab gels (Maxam & Gilbert, 1977) at 150 V (regulated) in a 4°C cold room. When the labeled DNA fragments required a secondary restriction digest before electrophoresis the denaturation step was omitted. Regions of the gel containing the desired

† Abbreviations used: kb, kilobases (10^3 bases); C_0t, moles nucleotide liter^{-1} second.

DNA fragments were localized by autoradiography and excised. The DNA was eluted by incubating crushed gel pieces in 1 ml 0·1 M-NaCl, 1 mM-EDTA, 10 mM-Tris·HCl (pH 7·4) overnight at 37°C. Acrylamide was removed by glass wool filtration, after which the DNA was bound to a small (~0·2 ml) bed of DE-52 (Whatman), prepared in a Pasteur pipette in the above buffer. The same buffer was used to wash the bound DNA, which was then eluted by application of 1·0 M-NaCl, 1 mM-EDTA, 0·1 M-Tris·HCl (pH 9·5) to the column. The DNA was precipitated with isopropanol and dissolved in distilled water.

(b) *DNA sequencing*

DNA sequencing was performed by the chemical degradation methods of Maxam & Gilbert (1977). Five cleavage reactions described by these authors were employed: guanine cleavage (G), strong guanine/weak adenine (G > A), strong adenine/weak cytosine (A+C), cytosine (C), and thymine/cytosine (C+T). Cleavage products were separated by electrophoresis on 7 M-urea/Tris/borate/EDTA/polyacrylamide gels (Maxam & Gilbert, 1977) at 1000 to 1500 V. Gels (40 cm long) of 3 different acrylamide concentrations (25%, 20% and 10%) were used routinely, in order to obtain optimal resolution for different (but overlapping) regions of the DNA sequence. Gels (25% and 20%) were poured with tapered spacers (0·5 mm at the top to 2 mm at the bottom (Schaffner *et al.*, 1978)) separating the glass plates; 10% gels were prepared with spacers of uniform 0·5 mm thickness (Sanger & Coulson, 1978). Autoradiography of sequencing gels was carried out at −80°C using preflashed Kodak XR-5 X-ray film and DuPont Lightning Plus intensifying screens (Laskey & Mills, 1977).

Sequencing was carried out on both strands of all strand-separated DNA fragments. This provided confirmation for most or all of each sequence. In addition, care was taken when reading the sequencing gel autoradiograms to detect *Eco*RII sites containing methylated Cs (Ohmori *et al.*, 1978).

(c) *Computer analysis of DNA sequences*

DNA sequences were analyzed using a PDP-11/34 computer (Digital Equipment Corp.). A set of programs for this purpose was developed by R. F. Murphy and J. W. Posakony (unpublished results), with the following capabilities: sequence entry; sequence complementing; rapid searching for restriction enzyme sites and other short sequence homologies; mono-, di-, and trinucleotide frequency determination; and general sequence comparison (using the matrix algorithm of Needleman & Wunsch (1970)).

3. Results

(a) *Nucleotide sequences of eight repetitive elements from the sea urchin genome*

The cloned DNA fragments used for sequencing were derived from renatured repeat duplexes, and all have been demonstrated to contain repetitive sequences in renaturation reactions with genomic DNA (Klein *et al.*, 1978). The approach used in constructing these clones was as follows (Scheller *et al.*, 1977*a*). Two thousand nucleotide long sea urchin DNA was renatured to C_0t 40 and the resulting structures were digested with S_1 nuclease. The duplex DNA fragments resisting digestion were ligated to synthetic decanucleotide "linkers" (Scheller *et al.*, 1977*b*) containing the recognition site for *Eco*RI restriction endonuclease. The ligated structures were digested with *Eco*RI, then joined to the plasmid vector RSF2124 and used to transform *Escherichia coli* strain C600. Since the individual members of sea urchin repetitive sequence families are divergent to various degrees (Klein *et al.*,

1978), the repetitive DNA duplexes that entered the host bacteria during transformation are likely to have been mismatched. Repair or replication in the host would be expected to correct such mismatches. Thus the sequences reported here could represent intermediates between the sequences of the two repeat family members which constituted the original DNA duplex. However, some of the families to which these sequences belong display very little internal sequence divergence, and in addition the S_1 nuclease treatment would have eliminated any duplexes with large non-base-paired regions. In any case the cloned sequences can be regarded as representative of their genomic repeat families. This has been demonstrated in a variety of ways, including thermal stability measurements on duplexes between the cloned fragments and homologous genomic DNA sequences (Klein *et al.*, 1978; Scheller *et al.*, 1981), as well as by the direct sequence comparisons given by Scheller *et al.* (1981) and in this paper.

Eight repeat clones were selected for DNA sequence analysis. Table 1 shows that

TABLE 1

Characteristics of eight cloned repetitive sequence elements

Clone	Length[a]	Reiteration frequency[b]	Divergence class[c]
CSp2034	498	2500	I
CSp2090	167	≥140	III
CSp2096	203	80	II[d]
CSp2108	204	20	II[d]
CSp2109 A	180	≥900	III
CSp2109 B	110	≥1000	III
CSp2111	155	12,500	—
CSp2112	257	450	II
CSp2137	226	530	II[d]

[a] Single-stranded length(s) (in nucleotides) of *Eco*RI sea urchin DNA fragment(s) in each clone, including synthetic linker sites, were determined from the DNA sequences.

[b] Approximate number of homologous sequences (repeat family members) per haploid *S. purpuratus* genome, determined by kinetics of reassociation with whole sea urchin DNA, in the presence of an internal single copy standard (Klein *et al.*, 1978). Data for CSp2034 and CSp2109A and B are from Anderson *et al.* (1981). The value for CSp2096 is from Moore *et al.* (1981).

[c] These classifications are based on the thermal stability of duplexes between the clone sequence and whole genomic DNA; they are a measure of the degree of sequence divergence within the repeat family represented by each clone. Δt_c is the difference between the mean thermal stability of the renatured probe DNA and of the population of heteroduplexes formed between the cloned repeat probe and the genomic DNA of *S. purpuratus*. Class I repeat families are those for which Δt_c is less than 4 deg.C. These families exhibit very little intrafamilial sequence divergence. Note that due to DNA sequence polymorphism, heteroduplexes formed by the single copy DNA of any 2 average *S. purpuratus* genomes melts about 4 deg.C below the t_m of perfect duplexes of the same length (Britten *et al.*, 1978). Class II families are those displaying moderate intrafamilial sequence divergence. They are defined as families for which Δt_c is more than 4 deg.C and for which lowering the criterion of the renaturation reaction from 55°C to 45°C in 0·12 M-sodium phosphate buffer (pH 6·8) does not reveal the presence of additional, more distantly related family members. Class III families are those for which a lowered criterion condition results in the appearance of additional reactive sequences in the genome, whose degree of homology is now sufficient to promote stable duplex formation. Measurements of reiteration frequency for class III families are therefore criterion-limited and yield only minimum estimates of family size. Δt_c for class III families is generally 20 deg.C or greater.

[d] The thermal stabilities of duplexes formed between this clone and whole genomic DNA appear to fall into 2 groups, representing relatively discrete classes of closely and distantly related sequences.

the set of repetitive sequence families represented by these clones includes families whose size differs by two orders of magnitude and families displaying both small and large degrees of internal sequence divergence. In Figure 1 are displayed the nucleotide sequences of the sea urchin DNA inserts in the eight clones. The synthetic *Eco*RI linkers used in the construction of these clones yield, upon cleavage, the sequence

5′-A-A-T-T-C-G-G-3′
3′- G-C-C-5′

(Scheller *et al.*, 1977*b*). As shown in Figure 1, the 5′ ends of most of the repeat fragments appear by their agreement with this sequence to have derived from linker sites, though in three of the clones (CSp2034, CSp2112 and CSp2137) one 5′ end apparently derived from a natural *Eco*RI site. Two sea urchin DNA fragments, A and B, are released from CSp2109 on digestion with *Eco*RI. Evidence discussed below indicates that the non-linker *Eco*RI site separating the CSp2109 A and B fragments was created by mismatch repair during cloning in the host bacterium.

Table 2 shows that most of the clones exhibit a G+C content somewhat higher than that of whole *Strongylocentrotus purpuratus* DNA. A calculation using the lengths and repetition frequencies of the sequenced clones (from Table 1) yields an approximate weighted average of 43% G+C for these sequence families compared to 36·6% for the *S. purpuratus* genome as a whole (Thiery *et al.*, 1976).

A "nearest neighbor" analysis of the sequences was carried out by determining the observed frequencies of the 16 possible dinucleotides, and comparing these values with those predicted from the products of the appropriate mononucleotide frequencies. Statistically significant ($P < 0{\cdot}01$) deviations from random expectation were found for some doublets in individual clones (data not shown). However, significant trends involving all of the clones and contrasting with measurements on total sea urchin DNA (Swartz *et al.*, 1962) were not observed.

(b) *Features of internal sequence organization*

The internal organization of the cloned repeat sequence elements was examined by searching each sequence with a computer for direct or inverted repetitions. Five of the clones (CSp2034, CSp2108, CSp2109, CSp2111 and CSp2112) contain

---

FIG. 1. Nucleotide sequences of individual sea urchin repetitive sequence elements cloned in the plasmid RSF2124 (Scheller *et al.*, 1977*a*) using synthetic *Eco*RI linkers (Scheller *et al.*, 1977*b*). Each sequence terminates with the site cut by the *Eco*RI endonuclease. Note, however, that *Eco*RI digestion of clone CSp2109 releases 2 sea urchin DNA fragments, A and B. The upper and lower lines of each sequence correspond, respectively, to the upper (slower) and lower (faster) strands of each DNA fragment when denatured and subjected to electrophoresis on neutral polyacrylamide strand separating gels. Numbers refer to the number of nucleotides from the 5′ end of the upper strand. A short sequence homology widely shared among the clones (see the text and Table 4) is enclosed by thin-lined boxes. Statistically significant direct or inverse sequence repetitions within given sequences are enclosed in thick-lined boxes (see Fig. 2). Vertical lines within a box separate the elements of a tandem direct repeat and vertical lines outside a box indicate the axis of a tandem inverse repeat. Sequences were determined by the method of Maxam & Gilbert (1977) (see Materials and Methods). Hyphens omitted for clarity (also from subsequent Figures).

CSp2034

```
1       10        20        30        40        50        60        70        80        90        100       110
AATTCCATTTCCCGTTTACGTGTAATCGCCATGGGAAATGTGTTGGTATATCGGTAACGCACTTGACTTCCAAGCTCGAGATACCTGGTTTGAGTCCGGGTGTCTCCAAC
    GGTAAAGGGCAAATGCACATTAGCGGTACCCTTTACACAACCATATAGCCATTGCGTGAACTGAAGGTTCGAGCTCTATGGACCAAACTCAGGCCCACAGAGGTTG

        120       130       140       150       160       170       180       190       200       210       220
AATTTTTGTCTTCATTTCTTTTTTCTTTTTGACCTTTCACTGTACCATCTCCTTTGATTATTAAAACATTATTCGCAAATAATGTAATAAACAACCGTATTTACCTTATT
TTAAAAACAGAAGTAAAGAAAAAAGAAAAACTGGAAAGTGACATGGTAGAGGAAACTAATAATTTTGTAATAAGCGTTTATTACATTATTTGTTGGCATAAATGGAATAA

        230       240       250       260       270       280       290       300       310       320       330
ATTACCCTTTTTTCCACTTAATCCATTATTTAGTGCGTGTAGTCCCTATGGGGGATGACTTTTTTTTCAAATCTAGATTCTACATTCCTGTCTTTTGCGAGCAATATCGC
TAATGGGAAAAAAGGTGAATTAGGTAATAAATCACGCACATCAGGGATACCCCCTACTGAAAAAAAAGTTTAGATCTAAGATGTAAGGACAGAAAACGCTCGTTATAGCG

        340       350       360       370       380       390       400       410       420       430       440
CGTTATTTCTGCTTGGATTTAGTTGATTTTTGAGTATGTTGTAGCTGAGACTATATGCTATCTAAATCATTCCTCATTTTTTCAATTAGATATCAGGATCACGCTTATGA
GCAATAAAGACGAACCTAAATCAACTAAAAACTCATACAACATCGACTCTGATATACGATAGATTTAGTAAGGAGTAAAAAAGTTAATCTATAGTCCTAGTGCGAATACT

        450       460       470       480       490       500
TCATGATCATTATCATCTCATTTCTCACTGAGTATTATTTATATGTCTGTCTCTCCCG
AGTACTAGTAATAGTAGAGTAAAGAGTGACTCATAATAAATATACAGACAGAGAGGGCTTAA
```

FIG. 1. (PART I)

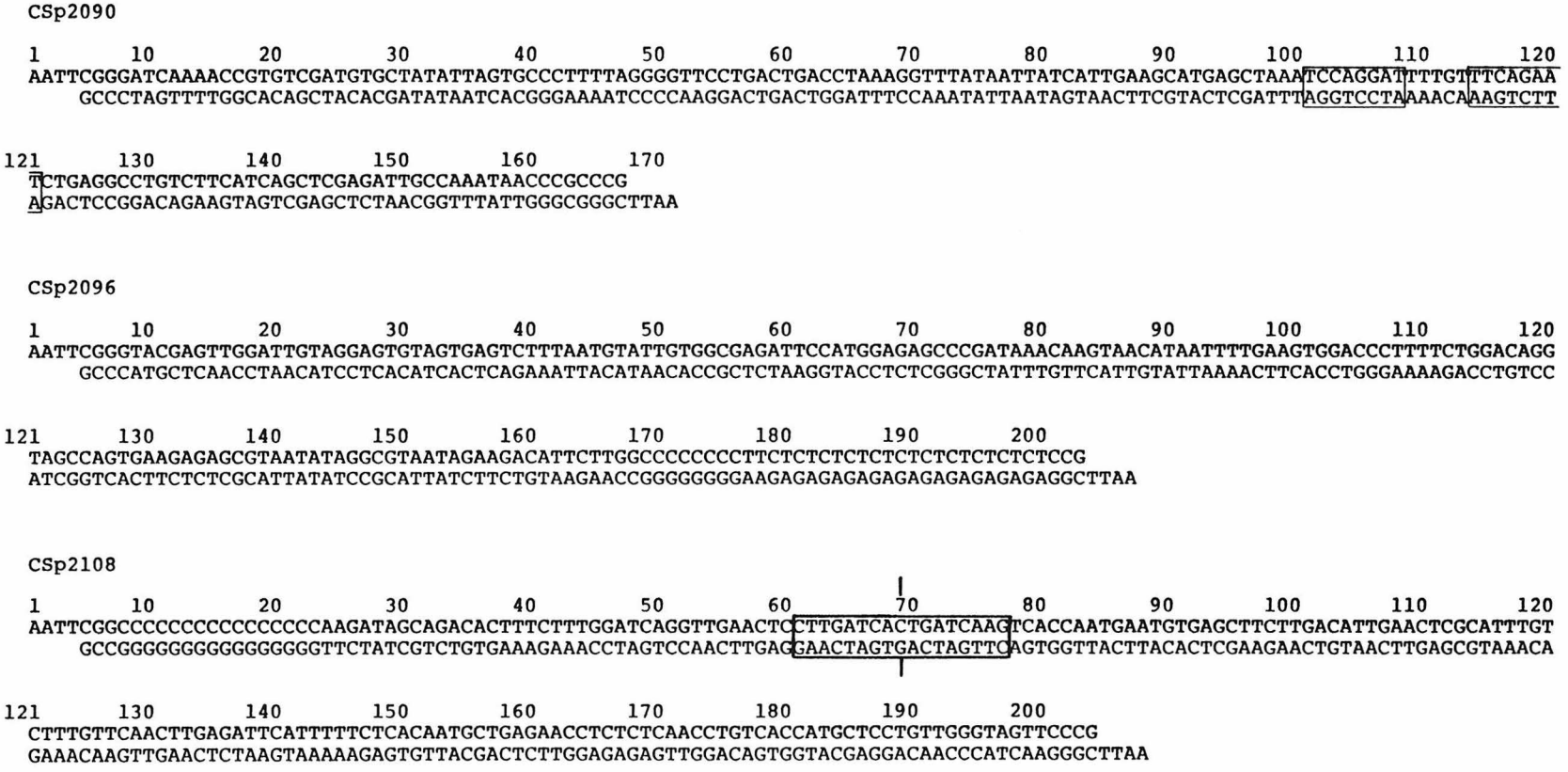

CSp2090
1 10 20 30 40 50 60 70 80 90 100 110 120
AATTCGGGATCAAAACCGTGTCGATGTGCTATATTAGTGCCCTTTTAGGGGTTCCTGACTGACCTAAAGGTTTATAATTATCATTGAAGCATGAGCTAAATCCAGGATTTTGTTTCAGAA
GCCCTAGTTTTGGCACAGCTACACGATATAATCACGGGAAAATCCCCAAGGACTGACTGGATTTCCAAATATTAATAGTAACTTCGTACTCGATTTAGGTCCTAAAACAAAGTCTT
121 130 140 150 160 170
TCTGAGGCCTGTCTTCATCAGCTCGAGATTGCCAAATAACCCGCCCG
AGACTCCGGACAGAAGTAGTCGAGCTCTAACGGTTTATTGGGCGGGCTTAA
CSp2096
1 10 20 30 40 50 60 70 80 90 100 110 120
AATTCGGGTACGAGTTGGATTGTAGGAGTGTAGTGAGTCTTTAATGTATTGTGGCGAGATTCCATGGAGAGCCCGATAAACAAGTAACATAATTTTGAAGTGGACCCTTTTCTGGACAGG
GCCCATGCTCAACCTAACATCCTCACATCACTCAGAAATTACATAACACCGCTCTAAGGTACCTCTCGGGCTATTTGTTCATTGTATTAAAACTTCACCTGGGAAAAGACCTGTCC
121 130 140 150 160 170 180 190 200
TAGCCAGTGAAGAGAGCGTAATATAGGCGTAATAGAAGACATTCTTGGCCCCCCCCTTCTCTCTCTCTCTCTCTCTCTCCG
ATCGGTCACTTCTCTCGCATTATATCCGCATTATCTTCTGTAAGAACCGGGGGGGGAAGAGAGAGAGAGAGAGAGAGAGGCTTAA
CSp2108
1 10 20 30 40 50 60 70 80 90 100 110 120
AATTCGGCCCCCCCCCCCCCCCCAAGATAGCAGACACTTTCTTTGGATCAGGTTGAACTCCTTGATCACTGATCAAGTCACCAATGAATGTGAGCTTCTTGACATTGAACTCGCATTTGT
GCCGGGGGGGGGGGGGGGGGTTCTATCGTCTGTGAAAGAAACCTAGTCCAACTTGAGGAACTAGTGACTAGTTCAGTGGTTACTTACACTCGAAGAACTGTAACTTGAGCGTAAACA
121 130 140 150 160 170 180 190 200
CTTTGTTCAACTTGAGATTCATTTTTCTCACAATGCTGAGAACCTCTCTCAACCTGTCACCATGCTCCTGTTGGGTAGTTCCCG
GAAACAAGTTGAACTCTAAGTAAAAAGAGTGTTACGACTCTTGGAGAGAGTTGGACAGTGGTACGAGGACAACCCATCAAGGGCTTAA

FIG. 1. (PART II)

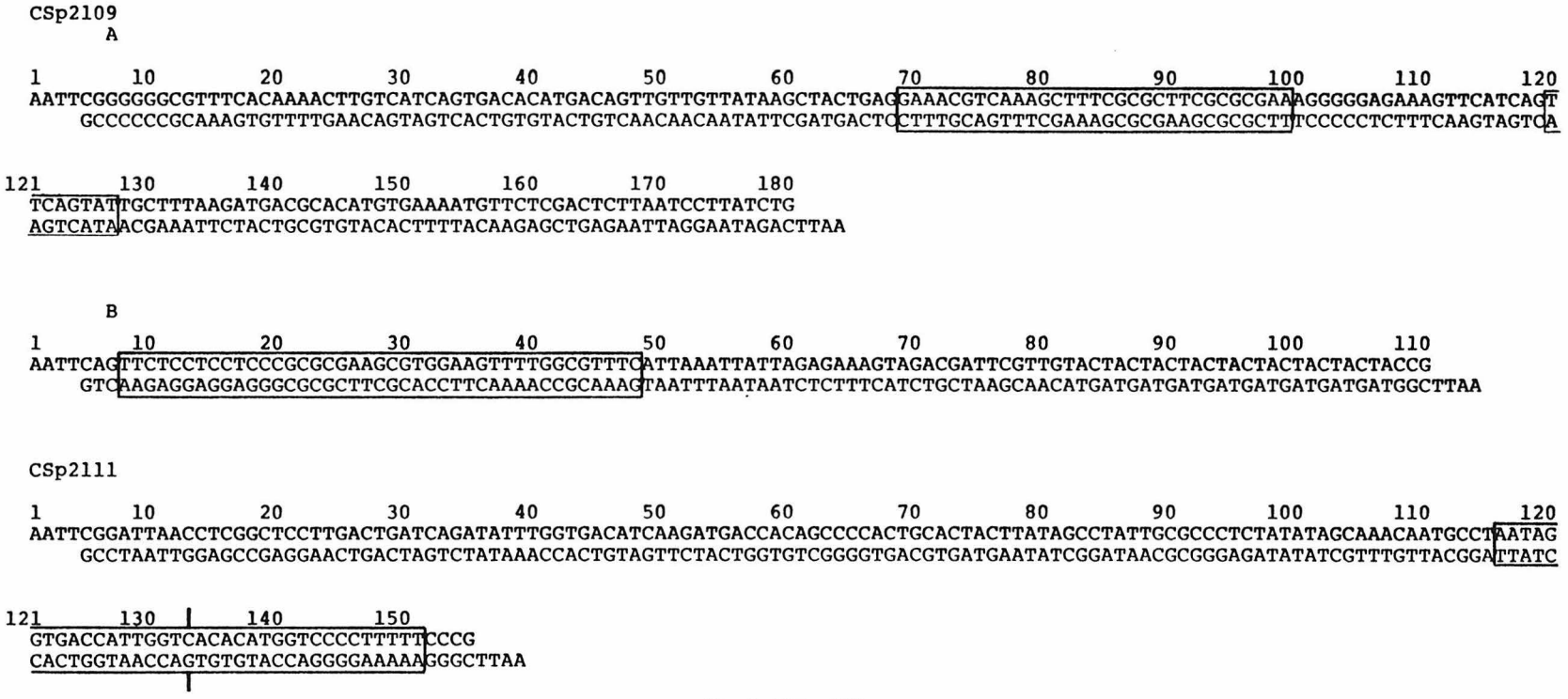


FIG. 1. (PART III)

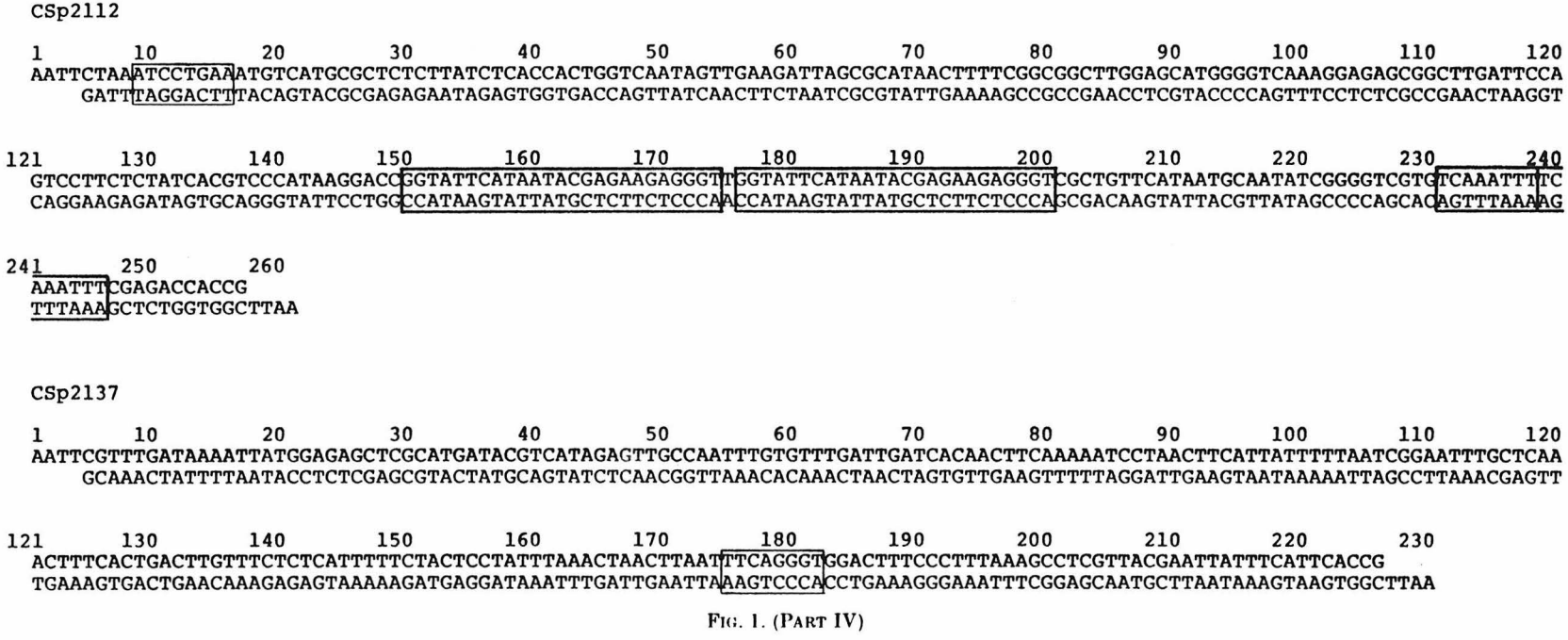

CSp2112
1 10 20 30 40 50 60 70 80 90 100 110 120
AATTCTAAATCCTGAAATGTCATGCGCTCTCTTATCTCACCACTGGTCAATAGTTGAAGATTAGCGCATAACTTTTCGGCGGCTTGGAGCATGGGGTCAAAGGAGAGCGGCTTGATTCCA
GATTTAGGACTTTACAGTACGCGAGAGAATAGAGTGGTGACCAGTTATCAACTTCTAATCGCGTATTGAAAAGCCGCCGAACCTCGTACCCCAGTTTCCTCTCGCCGAACTAAGGT
121 130 140 150 160 170 180 190 200 210 220 230 240
GTCCTTCTCTATCACGTCCCATAAGGACCGGTATTCATAATACGAGAAGAGGGTTGGTATTCATAATACGAGAAGAGGGTCGCTGTTCATAATGCAATATCGGGGTCGTGTCAAATTTTC
CAGGAAGAGATAGTGCAGGGTATTCCTGGCCATAAGTATTATGCTCTTCTCCCAACCATAAGTATTATGCTCTTCTCCCAGCGACAAGTATTACGTTATAGCCCCAGCACAGTTTAAAAG
241 250 260
AAATTTCGAGACCACCG
TTTAAAGCTCTGGTGGCTTAA
CSp2137
1 10 20 30 40 50 60 70 80 90 100 110 120
AATTCGTTTGATAAAATTATGGAGAGCTCGCATGATACGTCATAGAGTTGCCAATTTGTGTTTGATTGATCACAACTTCAAAAATCCTAACTTCATTATTTTTAATCGGAATTTGCTCAA
GCAAACTATTTTAATACCTCTCGAGCGTACTATGCAGTATCTCAACGGTTAAACACAAACTAACTAGTGTTGAAGTTTTTAGGATTGAAGTAATAAAAATTAGCCTTAAACGAGTT
121 130 140 150 160 170 180 190 200 210 220 230
ACTTTCACTGACTTGTTTCTCTCATTTTTCTACTCCTATTTAAACTAACTTAATTTCAGGGTGGACTTTCCCTTTAAAGCCTCGTTACGAATTATTTCATTCACCG
TGAAAGTGACTGAACAAAGAGAGTAAAAAGATGAGGATAAATTTGATTGAATTAAAGTCCCACCTGAAAGGGAAATTTCGGAGCAATGCTTAATAAAGTAAGTGGCTTAA

FIG. 1. (PART IV)

TABLE 2

G+C content of repetitive DNA sequences

Clone	G+C content (%)[a]
CSp2034	35
CSp2090	42
CSp2096	45
CSp2108	46
CSp2109A	43
CSp2109B	40
CSp2111	46
CSp2112	44
CSp2137	32
Whole genome DNA[b]	36·6

[a] Determined from the DNA sequences; excludes synthetic *Eco*RI linker sites.

[b] Thiery *et al.* (1976).

statistically significant features of this kind ($P < 0{\cdot}01$; Dykes *et al.*, 1975), and three do not. Examples were found of both direct and inverted repetitions in which the repeated subelements occurred either adjacent to each other or separated in the sequence. Figure 2 shows block diagrams in which the location and extent of statistically significant internal repetitions can be observed in the five sequences that contain them. A peculiar and statistically improbable feature observed in four of the eight clones is the presence of short nucleotide sequences containing both direct and inverted repetitions (the relevant sequences are boxed and can be examined in Fig. 1). It should be noted that the internal repetition shown in Figure 2 for CSp2108 is found unaltered in a homologous cloned fragment of genomic DNA isolated from a λ recombinant (λ2108-16A; sequence presented in Fig. 5(e) of Scheller *et al.* (1981)).

Though the features indicated in Figure 2 are all very unlikely to have occurred randomly (see legend), it is important to note that only a minor fraction of the length of repetitive sequence included in the cloned repeats is occupied by these features. The longest regions of internal sequence repetition were found in clones CSp2112 and CSp2109 and these are reproduced in Figure 3(a) and (b). As diagrammed in Figure 3(a), CSp2112 contains a 68 nucleotide region of direct repetition that includes an alternating tandem repeat of the form ϛβϛβϛ which can also be considered a 25 nucleotide exact repeat spaced by one nucleotide. Figure 3(b) shows that the sequences of the A and B fragments from CSp2109 share a 41 nucleotide region of homology. This observation suggests that these two fragments were not cloned together by chance (i.e. *via* ligation of unrelated fragments) but instead resulted from the cloning of a single DNA duplex which then acquired an internal *Eco*RI site by mismatch repair in the host bacterium. This site is probably not a natural one since in the construction of the repeat clones an *Eco*RI digestion was carried out after ligation of the linkers to the S_1 nuclease-

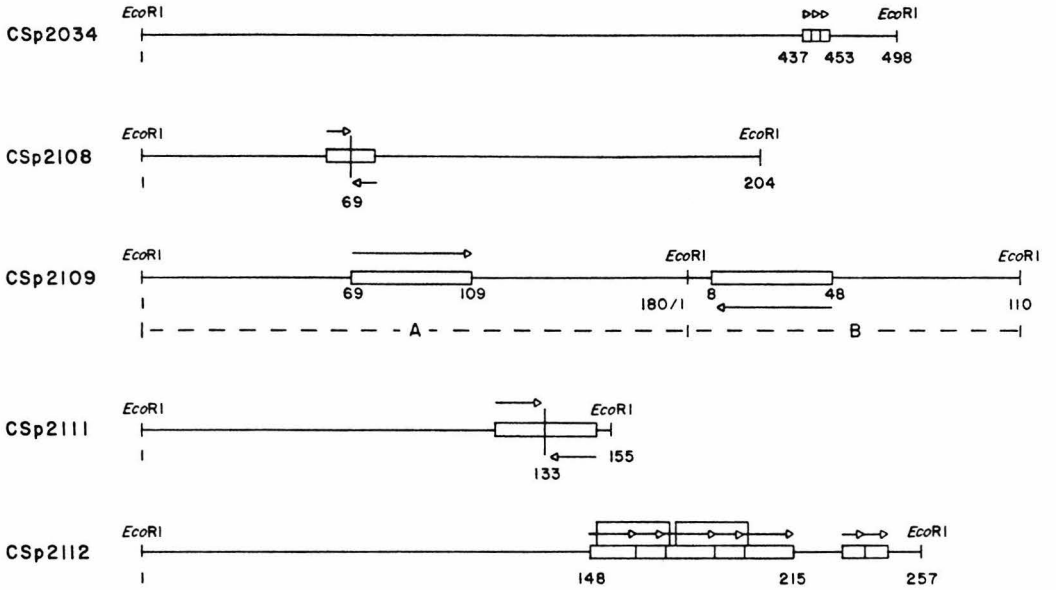


FIG. 2. Block diagrams of statistically significant sequence repetitions found within 5 of the 8 sequenced repeat elements (Fig. 1). Solid horizontal lines show the length of each sequence to scale (the scale for CSp2034 is one-half that for the other sequences). Boxes represent the location and extent of the internally repeated sequence elements. Vertical lines within boxes separate the elements of tandem repetitions. Horizontal arrows indicate the relative 5′ to 3′ orientations of the elements. Numbers shown correspond to those in Fig. 1. The probabilities that the repetitions illustrated occurred randomly were calculated essentially as by Dykes *et al.* (1975). CSp2034: 6/6, 6/6, 5/6 tandem direct repetition, $P = 0{\cdot}002$. CSp2108: 8/8 tandem inverse repetition, $P = 0{\cdot}006$. CSp2109: 34/41 separated inverse repetition, $P = 2 \times 10^{-10}$; within this larger repetition is a 9/10 tandem inverse repetition (see Fig. 3), $P = 0{\cdot}015$. The sequences of the A and B fragments of CSp2109 are shown forming a continuous repetitive element, corresponding to their genomic relationship (see the text). CSp2111: 13/17 tandem inverse repetition, $P = 0{\cdot}003$. CSp2112: alternating tandem direct repetition in which 3 direct repeats are separated by 2 other direct repeats. This region can also be considered a 25/25 direct repeat separated by 1 nucleotide, $P = 4 \times 10^{-13}$. CSp2112 also contains an 8/8 tandem direct repeat, $P = 0{\cdot}007$. No statistically significant internal sequence repetitions were found in CSp2090, CSp2096 or CSp2137, and these sequences (see Fig. 1) are hence omitted from this Figure.

resistant repetitive DNA duplexes (Scheller *et al.*, 1977*a*). As noted above, the A and B fragments of CSp2109 each have one linker type and one non-linker type 5′-terminal sequence (Fig. 1). Further evidence comes from the sequence of a cloned genomic fragment discussed below (λ2109B-18) which bridges the position of the internal *Eco*RI site in CSp2109 and which includes a continuous sequence homologous to the A and B portions of the CSp2109 sequence in the order shown in Figure 3(b). Thus CSp2109 contains a naturally occurring 41 nucleotide inverse, separated repetition including 34 matched nucleotides. This is the longest inverse repeat discovered in any of the eight sequenced clones. However, this mismatched inverse repeat could not form an intrastrand DNA duplex which would be stable under standard criterion conditions (60°C, 0·18 M-Na^+), and the same is true of all the shorter inverse repeats indicated in Figure 2.

Three of the eight clones were found to contain stretches of simple sequence repetition located immediately adjacent to a linker-type 5′ end (Fig. 1): CSp2096,

$(AG)_{11}$; CSp2108, $(C)_{16}$; CSp2109B, $(TAG)_8$. However, the sequence of a cloned genomic DNA fragment homologous to CSp2108 (λ2108-16A; see Fig. 5(e) of Scheller *et al.*, 1981) does not display this run of Cs in the corresponding location, although the λ2108-16A sequence is otherwise only 4% different from the CSp2108 sequence. It is possible that these simple terminal sequence repetitions are not of genomic origin. On the other hand, some genomic members of the repeat families could contain such simple sequence elements, and the cloned examples could be representative of such members.

(c) *Translation termination signals in repeat sequences*

Previous work demonstrated that transcripts complementary to the great majority of repetitive sequence families in the sea urchin genome are present in the RNA of the sea urchin egg (Costantini *et al.*, 1978; Moore *et al.*, 1980). This is true of all the eight sequenced repeats. Recently we found that most of the repeat transcripts are covalently linked to single copy sequences in egg poly(A)$^+$ RNA (Costantini *et al.*, 1980). Although there are various reasons to doubt that a protein coding function exists for the repeat portion of the maternal poly(A) RNAs (see Discussion), we examined the eight repeat sequences for the presence of translation termination codons. Table 3 shows that stop codons are found in all three phases of both strands of six out of eight of the clones sequenced. In one of the remaining

TABLE 3

Translation termination signals in repetitive sequence elements

Clone and strand[a]	Number of stop codons in phase[b]		
	1	2	3
CSp2034 U	11	8	6
CSp2034 L	18	10	9
CSp2090 U	3	4	4
CSp2090 L	1	4	7
CSp2096 U	6	5	3
CSp2096 L	1	0	0
CSp2108 U	2	9	0
CSp2108 L	4	0	5
CSp2109 U	5	1	6
CSp2109 L	4	5	16
CSp2111 U	2	5	3
CSp2111 L	1	2	6
CSp2112 U	4	4	3
CSp2112 L	6	4	4
CSp2137 U	7	5	1
CSp2137 L	6	9	7

[a] The strand designations upper (U) and lower (L) correspond to the relative positions of the strands on neutral polyacrylamide strand separation gels (see Materials and Methods), and also to the upper and lower sequence lines of Fig. 1.

[b] Determined from the DNA sequences (Fig. 1); the phases are arbitrary.

sequences, CSp2108, a very closely related genomic sequence belonging to the same repeat sequence subfamily has all reading frames blocked by termination codons (Scheller *et al.*, 1981). We conclude that the sea urchin repeat sequences are not usually translatable. However, it is important to note that since each family is more or less divergent (Table 1), it cannot be excluded that there are certain members of given repeat families which preserve one or more open reading frames.

(d) *Sequence relationships between different repeat families*

To examine the possibility that different repeat families may be related by short homologies we carried out detailed pairwise comparisons of the cloned sequences. The most important result is that no statistically significant homology shared by all or many of the cloned repeats could be detected. That is, the repetitive sequence families represented by these clones are not in general related above random expectation. One interesting homology confined to two of the repeat clones is shown in Figure 4. As indicated there, a continuous 44 nucleotide block of the CSp2108

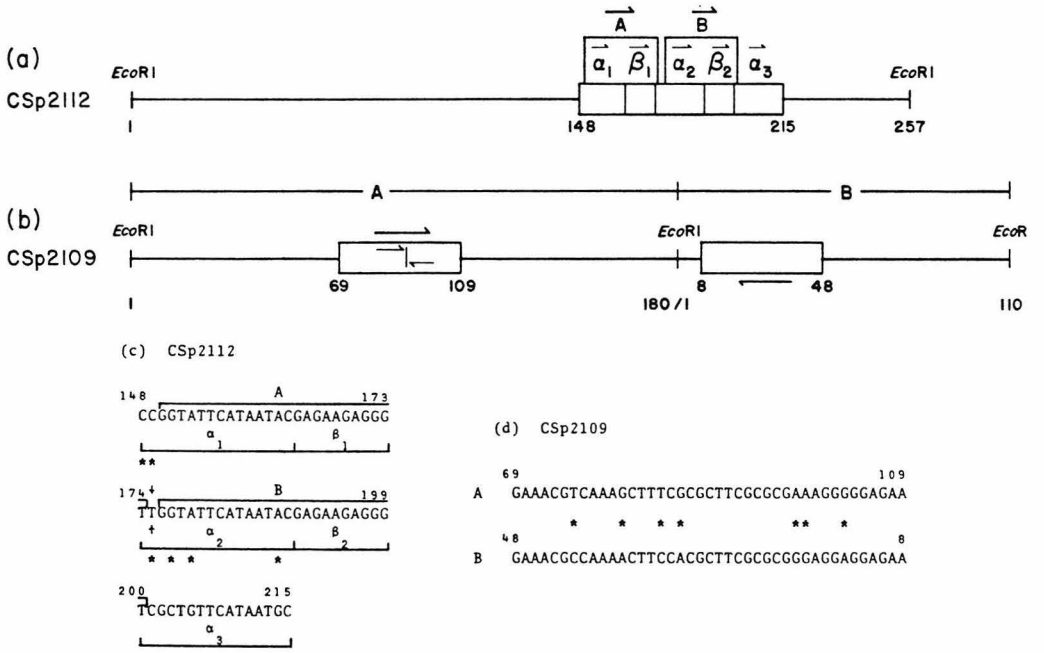


FIG. 3. Detail of the major sequence element repetitions in CSp2112 ((a) and (c)) and CSp2109 ((b) and (d)). (a) and (b) Block diagrams. Solid horizontal lines show the length of each sequence to scale. Boxes represent the location and extent of repeated sequence elements. Vertical lines within boxes separate the elements of tandem repetitions. Horizontal arrows indicate the relative 5′ to 3′ orientations of the elements. Numbers shown correspond to those of Fig. 1. In CSp2112 (a), 3 direct repeats (α_1, α_2, α_3) are separated by 2 other direct repeats (β_1, β_2). This sequence region also includes a 25/25 direct repeat (A, B) separated by 1 nucleotide. In CSp2109 (b), the large repeat element in *Eco*RI fragment A itself contains a 9/10 tandem inverse repetition as shown. (c) and (d) Alignments of the sequences of the repetitions diagrammed in (a) and (b), respectively. Numbers above the sequences correspond to those of Fig. 1, and sequence elements in (c) are labeled with the designations shown in (a). Asterisks indicate non-homologous bases.

sequence is found in CSp2137, divided (in order) into four nearby though separate segments. This homology would not be expected to occur by chance.

We also investigated the extent to which the eight sequences might display simple underlying repeating sequence structures. The frequencies of occurrence of all exact internal repetitions of five nucleotides or longer were measured. A detailed comparison showed that the frequency of occurrence of short internal repeat features is no greater than is found in the rabbit β globin gene sequence (Hardison *et al.*, 1979) or in randomly generated nucleotide sequences. The interspersed repetitive elements in sea urchin DNA clearly do not consist of tandem or of frequent repetitions of very short sequence subelements. Their lack of either extensive internal homology or significant intersequence homology means that the *sequence complexity* of the eight cloned repeats is approximately equal to the sum of the lengths of the cloned inserts. Thus if the clones we have sequenced are representative, the rate of reaction of the repetitive DNA sequences in the sea urchin genome observed in renaturation studies should provide a reasonable approximate estimate of the number of different repeat sequence families in this genome. Kinetic determinations carried out previously on sea urchin DNA (Graham *et al.*, 1974; Klein *et al.*, 1978; Costantini *et al.*, 1978,1980) suggested that there are at least several thousand interspersed repeat families in the DNA of *S. purpuratus*.

A short sequence element shared by five of the eight clones was also detected (Table 4), but statistical calculations show that this homology could have arisen by chance. The consensus of the shared sequence is

5′-T-T-C-A-G-G-A-T-3′
3′-A-A-G-T-C-C-T-A-5′

TABLE 4

A sequence homology widely shared among eight repetitive elements

Clone	Sequence[a]
CSp2034	A-T-C-A-G-G-A-T
CSp2090	T-C-C-A-G-G-A-T
	T-T-C-A-G-A-A-T
CSp2096	—
CSp2108	—
CSp2109	T-T-C-A-G-T-A-T
CSp2111	—
CSp2112	T-T-C-A-G-G-A-T
CSp2137	T-T-C-A-G-G-G-T
CS clone consensus	T-T-C-A-G-G-A-T
Splice junction consensus[b]	
Left	A/C A-G-G-T-A
Right	Y-T-C-A-G-G

[a] Only sequences in which at least 7/8 nucleotides match with the 8 nucleotide consensus sequence are shown. All 8 clones contained 6/8 or better matches.

[b] Lerner *et al.* (1980). Sequences shown are for the 5′ (left) and 3′ (right) boundaries of intervening sequences. Y = pyrimidine.

A total of six examples of either this sequence or a 7/8 homologue was observed in the five clones. The location of this element in each of the repeat sequences is indicated in Figure 1. As pointed out earlier (Davidson & Britten, 1979) this sequence may be of interest because it is similar to that noticed by Breathnach *et al.* (1978) and Lerner *et al.* (1980) in junctional regions between translated messenger RNA sequences and the intervening sequences of various vertebrate and viral genes.

(e) *Sequence relationships between genomic members of the 2109 repeat family*

The repetitive sequence family represented by clone CSp2109 was chosen for further study for several reasons. The members of this family are generally short sequences interspersed with single copy DNA (Anderson *et al.*, 1981), a pattern typical of most sea urchin repeat families. In addition this family exhibits a substantial degree of intrafamilial sequence divergence. This was shown by measuring the thermal stability of heteroduplexes formed between CSp2109B DNA and whole sea urchin DNA (Klein *et al.*, 1978) or between CSp2109B DNA and individual λ recombinants containing homologous 2109 sequence elements (Scheller *et al.*, 1981). The sequence comparisons shown below provide specific examples of the relationships and distinctions which may exist among the members of a divergent interspersed repeat family.

Figure 5 presents the nucleotide sequences of two fragments of genomic DNA,

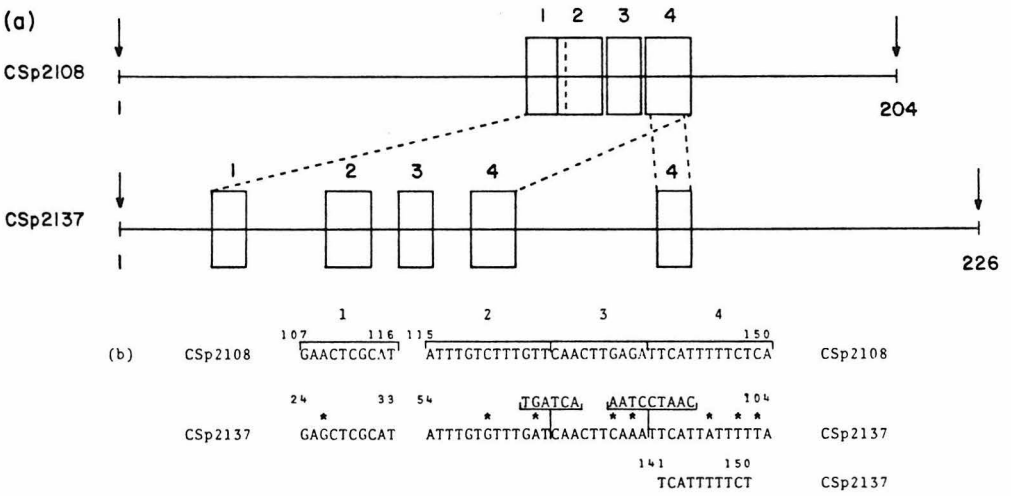


FIG. 4. Sequence homologies between CSp2108 and CSp2137. (a) Block diagram. Solid horizontal lines show the length of each sequence to scale. *Eco*RI sites are indicated by downward arrows and the numbers at each site correspond to those of Fig. 1. A continuous 44 nucleotide region in CSp2108 is divided (in order) into 4 separated segments (boxes labeled 1, 2, 3, 4) in CSp2137. A second homology in CSp2137 to segment 4 of CSp2108 is also shown. (b) Detail of the sequence relationships between CSp2108 and CSp2137 diagrammed in (a). Numbers above the sequences correspond to those of Fig. 1, and the sequence segments are labeled with the designations shown in (a). The numbered sequence elements of CSp2108 are shown on the upper line, and homologous segments of CSp2137 are aligned directly underneath. Asterisks mark non-homologous bases.

λ2109B-12

```
1       10        20        30        40        50        60        70        80        90        100       110       120
AATTCAGTTTCCCTCCCTCCCGCGCGATGCGCGGAAGTTTTGACGTTTCATTAAATTATCAGAGAAATTGTCTGAATTAATTTTTTCTTTTCTTTGTCCGATAAAAAAGTCACAGTTCGT
    GTCAAAGGGAGGGAGGGCGCGCTACGCGCCTTCAAAACTGCAAAGTAATTTAATAGTCTCTTTAACAGACTTAATTAAAAAAGAAAAGAAACAGGCTATTTTTTCAGTGTCAAGCA

121     130
AGGGAGAAAA . . .
TCCCTCTTTT . . .
```

λ2109B-18

```
1       10        20        30        40        50        60        70        80        90        100       110       120
TCGAGTGTATAAACTGTGTTGTTATCAGTGGCGTCGATCGCTCGACGTTTCATTAAATTATTAGAGAAAGAAGACGATTCGTTGTACTGTTTTCGACTCCTAATCCTTGTCTGAATAAAG
    CACATATTTGACACAACAATAGTCACCGCAGCTAGCGAGCTGCAAAGTAATTTAATAATCTCTTTCTTCTGCTAAGCAACATGACAAAAGCTGAGGATTAGGAACAGACTTATTTC

121     130       140       150       160       170       180       190
TTTCCCCCCTCCCGCGCGAAACGCGGAAGCTTTGGCGTTTCATTTGATTATAACTCTTAATCCTTGTTTG
AAAGGGGGGAGGGCGCGCTTTGCGCCTTCGAAACCGCAAAGTAAACTAATATTGAGAATTAGGAACAAACTTAA
```

FIG. 5.

each containing a member of the 2109B repeat family. The fragments were isolated from the recombinants λ2109B-12 and λ2109B-18, which had been selected from a genome library by reaction with the 2109B probe (see Anderson *et al.* (1981) for a restriction map and other data on these DNA fragments). Comparison of these sequences to those of the A and B fragments of the CSp2109 clone revealed the patterns of homology illustrated in Figure 6. Here it can be seen that λ2109B-18 contains a continuous 86 nucleotide segment (region II) homologous to a region that includes the internal *Eco*RI site in the plasmid clone CSp2109. Though in some locations in the genome the A and B sequences cloned in CSp2109 apparently occur together, this is not always the case. Thus some genomic λ recombinants isolated by Anderson *et al.* (1981) react with the 2109A probe but not with the 2109B probe. Furthermore, the developmental variation of 2109A transcript levels in sea urchin RNAs is different from that of 2109B transcripts (Scheller *et al.*, 1978). The 2109A and B repeat sequence families may have at least a partially separate physiological significance.

The most important observation in Figure 6 concerns the different arrangement of homologous sequence elements in λ2109B-18 and in CSp2109. These elements are shown as contiguous "blocks" whose order in CSp2109 is 1-2-3-4-5 (total sequence included in these blocks is 111 nucleotides). In the λ2109B-18 sequence the order is 4-5-1-2-3-4-2 (total length 147 nucleotides). There are several aspects of the remarkable alternative arrangement of sequence blocks which should be noted: (1) the boxed elements in the λ2109B-18 sequence are immediately adjacent to one another, and no other sequences intervene despite the change in order. (2) The various elements maintain the same orientation in the λ2109B-18 sequence as in CSp2109. (3) Blocks 4 and 2 are each repeated within the λ2109B-18 sequence. (4) At one terminus of the sequenced fragment is an *Eco*RI site (at the right end of the second occurrence of block 2) similar to that separating the A and B fragments of the CSp2109 clone.

Figure 6 shows that λ2109B-12 also includes a continuous region of homology with CSp2109, consisting of blocks 3 and 4 and a small part of block 5. The homology is delimited on the left by an *Eco*RI site, again analogous to the internal *Eco*RI site in CSp2109. On the right, however, the homology abruptly ceases and no further substantial homology occurs within the next 60 sequenced nucleotides or in other regions of the sea urchin DNA insert in λ2109B-12. Note that both λ2109B-18 and λ2109B-12 contain a 41 nucleotide sequence element that is repeated in inverted orientation in the A and B fragments of CSp2109.

A heteroduplex between two λ recombinants bearing different 2109B sequence elements is shown in Figure 7. As is usual for diverse members of the 2109 sequence

---

FIG. 5. Nucleotide sequences of 2 fragments of genomic DNA isolated from a λ recombinant library on the basis of their homology to the B probe sequence isolated from clone CSp2109 (Anderson *et al.*, 1981). The sequence labeled λ2109B-12 is the relevant region of an *Eco*RI + *Hin*fI restriction fragment. The sequence labeled λ2109B-18 is the complete sequence of an *Eco*RI + *Xho*I restriction fragment obtained from the clone λ2109B-18. No detectable homology to the 2109B probe sequence occurs outside the sequenced region of either clone. The sequences are oriented to correspond with the sequence of CSp2109B, as shown in Fig. 1.

family the DNA surrounding the complementary repeat sequences in these recombinants is non-homologous (cf. Anderson *et al.*, 1981). The interesting feature shown in Figure 7 (see insert, in particular) is that the two 2109B sequence elements differ by the presence (or absence) of a discrete sequence subelement, visualized as a loop in the heteroduplex region. This demonstrates that a subelement deletion (or insertion) has occurred leading to a difference in the sequence subelement content in the two repeats. This observation is of course on a larger scale than those of Figure 6, but its basic import may be similar. This is that members of the 2109 repeat family are not necessarily colinear and may differ in the arrangement and in the presence of particular sequence subelements.

(f) *Thermal stability of heteroduplexes between 2109 repeats*

The observations shown in Figure 6 require reinterpretation of the relatively low thermal stability of heteroduplexes formed between different members of the 2109 repeat family. Low thermal stability in repeat sequence duplexes has usually been considered the result of base-pair mismatch stemming from random sequence divergence. However, short duplexes also display lowered thermal stability, even if

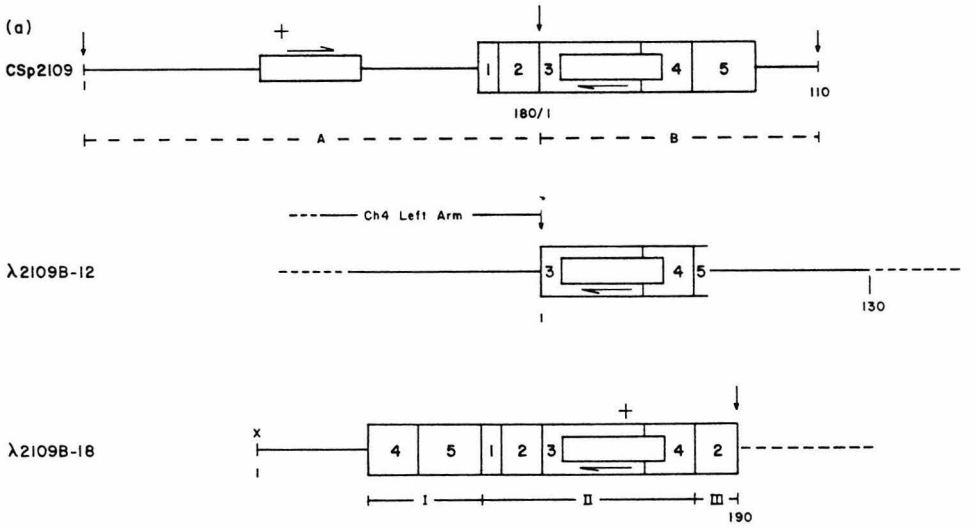


FIG. 6. Sequence relationships among CSp2109, λ2109B-12 and λ2109B-18. Solid horizontal lines show the lengths of the sequences to scale. (a) Block diagram. The numbers underneath each line correspond to those in Figs 1 and 5, and selected restriction sites are indicated: (↓) *Eco*RI; (+) *Hin*dIII; (×) *Xho*I. The large boxes enclose regions of homology among the 3 sequences and each region is labeled with a number (1 to 5). The 41 nucleotide inverse repeat in CSp2109 is represented by smaller boxes, and the corresponding regions of λ2109B-12 and λ2109B-18 are also shown, with arrows indicating relative 5′ to 3′ orientations. The diagram shows that the sequences in λ2109B-12 homologous to CSp2109 are immediately adjacent to the left arm of the Charon 4 cloning vector. Three regions of continuous homology to CSp2109 in λ2109B-18 are indicated by Roman numerals. (b) Sequence details of the numbered regions of homology. The top line shows the sequences of the 5 sequence blocks of CSp2109 shown in (a). Below this are shown the sequences of homologous elements in λ2109B-12 and λ2109B-18. In the case of λ2109B-18 regions of continuous homology with CSp2109 are indicated by Roman numerals, corresponding to those in (a). Asterisks indicate mismatched bases, relative to the CSp2109 sequence.

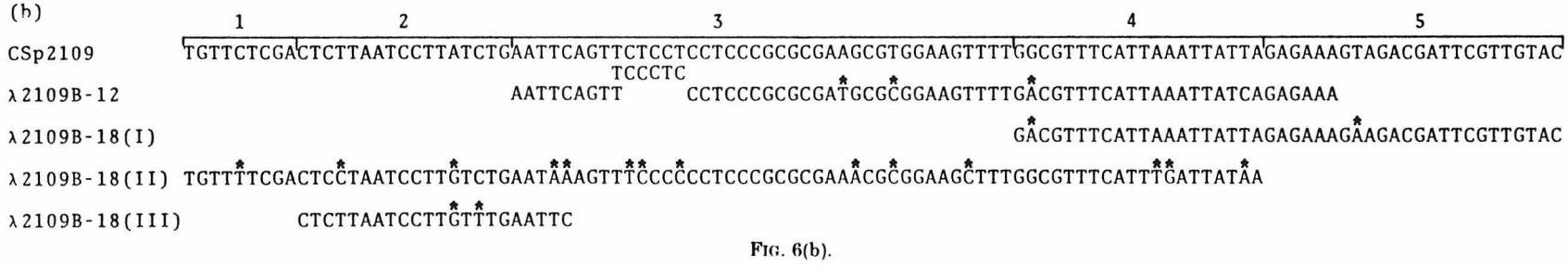

(b)
1
2
3
4
5
CSp2109 TGTTCTCGACTCTTAATCCTTATCTGAATTCAGTTCTCCTCCTCCCGCGCGAAGCGTGGAAGTTTTGGCGTTTCATTAAATTATTAGAGAAAGTAGACGATTCGTTGTAC
TCCCTC
λ2109B-12 AATTCAGTT CCTCCCGCGCGATGCGCGGAAGTTTTGACGTTTCATTAAATTATCAGAGAAA
λ2109B-18(I) GACGTTTCATTAAATTATTAGAGAAAGAAGACGATTCGTTGTAC
λ2109B-18(II) TGTTTTCGACTCCTAATCCTTGTCTGAATAAAGTTTCCCCCCTCCCGCGCGAAACGCGGAAGCTTTGGCGTTTCATTTGATTATAA
λ2109B-18(III) CTCTTAATCCTTGTTTGAATTC

FIG. 6(b).

perfectly matched. A consequence of the alternative sequence subelement orders shown in Figure 6 is that the continuous lengths of duplex which could form between the 2109 family members are short enough to affect significantly the thermal stability. Thus the longest continuous alignment in the CSp2109 and λ2109B-18 sequences is 86 nucleotides (ignoring base-pair mismatches); between CSp2109 and λ2109B-12 (which may not include a complete family member) only a 67 nucleotide segment is homologous; while λ2109B-12 and λ2109B-18 are homologous over only 60 nucleotides.

Scheller *et al.* (1981) found that the thermal stability of heteroduplexes formed between the B fragment of CSp2109 and the λ2109B-18 sequence shown in Figure 5 is 20 deg.C below that of the renatured 2109B DNA. Were this difference (Δt_m) due wholly to scattered base-pair mismatch it would imply a 20% difference between the two sequences (Britten *et al.*, 1974). These two sequences differ, however, by only 12% (calculated over the total homologous sequence length). For this case two alternative heteroduplexes could form, involving either region I or region II of the λ2109B-18 sequence (Fig. 6). The expected Δt_m for each of these alternatives is calculated in Table 5, taking into account the empirical relation between duplex length and duplex thermal stability (see legend). The net predicted Δt_m values for

TABLE 5

Calculated thermal stability of heteroduplexes between CSp2109B fragment and the genomic 2109B-18 sequence[a]

Factor affecting t_m	Homoduplex	Heteroduplex[b]	
	CSp2109B × CSp2109B	CSp2109B × 2109B-18 (I)	CSp2109B × 2109B-18 (II)
Base-pair mismatch[c] (%)	(0)	2·3	15·1
Δt_m (deg.C)	—	2·3	15·1
G+C content[d] (%)	43·4	30·2	43·4
Δt_m (deg.C)	—	5·4	0
Duplex length[e] (nucleotides)	106	43	53
Δt_m (deg.C)	—	9·0	6·1
Total Δt_m[f] (deg.C)	—	16·7	21·2

[a] See text for details. Values other than Δt_m were determined from the DNA sequences (Figs 1 and 5). Δt_m is the difference between the thermal stability of the heteroduplex and the thermal stability of the renatured B fragment probe from CSp2109 clone.

[b] Using the numbering shown in Figs 1 and 5, the following bases were considered to be involved in the heteroduplexes: CSp2109B (43–85)/λ2109B-18 (I) (46–88) and CSp2109B (6–58)/λ2109B-18 (II) (119–171).

[c] (Number of mismatched bases)/(number of mismatched bases + number of matched bases) × 100%. Δt_m = 1 deg.C/% base-pair mismatch (reviewed by Britten *et al.*, 1974; Wetmur, 1976).

[d] (Number of actual G+C base-pairs)/(total number of actual base-pairs) × 100%. (Mismatches were ignored in determining these values.) Δt_m = % (G+C)/2·44 (Mandel & Marmur, 1968).

[e] Total length of duplex region, including mismatches. $\Delta t_m = t_m(\infty) - t_m(L) = B/L$, where $t_m(\infty)$ is the t_m of an infinitely long DNA duplex, $t_m(L)$ is the t_m of a duplex of length L (in nucleotides) and B = 650. an empirical coefficient (Britten *et al.*, 1974). Thus here Δt_m = 650/(heteroduplex length) − 650/(homoduplex length).

[f] The 3 factors affecting t_m are assumed to act independently, so the total Δt_m is calculated as the sum of the 3 constituent Δt_m values.

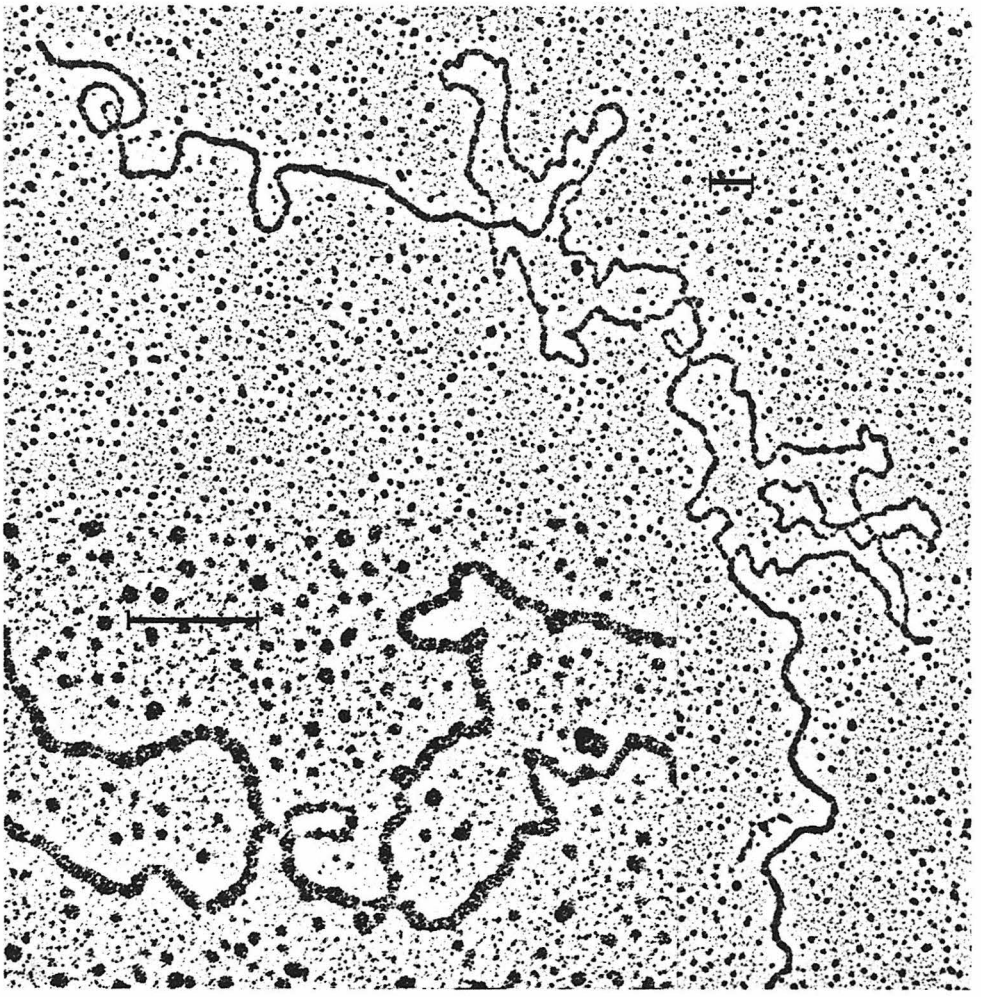


FIG. 7. Electron micrographs of a heteroduplex between 2 λ recombinants bearing 2109B sequences. (a) View of total duplex molecule formed by reaction of λ2109B-52 and λ2109B-55 (Anderson *et al.*, 1981). The sea urchin inserts begin at the forks marking the termini of the Charon 4 vector arms. One region of homology about 670 nucleotides in length exists within the sea urchin DNA inserts. This region contains the 2109B sequence (data not shown). As can be seen more clearly in the higher magnification insert, a 780 nucleotide sequence present in one of the genomic sequences is absent from the repeat sequence in the other. The structure shown is reproducible. The bar represents 0·5 kb.

region I and region II of λ2109B-18 are 17 deg.C and 21 deg.C, respectively, compared to the observed Δt_m of 20 deg.C. A similar analysis may be carried for λ2109B-12. We conclude that duplex length is indeed a major factor in the relative heteroduplex thermal stability decrease observed when 2109 family members are reacted together. The observations on the 2109 family show that the degree of sequence conservation in short regions of the repeat element is significantly greater than estimated by measuring overall heteroduplex thermal stability.

4. Discussion

(a) *General implications of sequence data*

There are several interpretations of repetitive sequences that are not consistent with the sequence information presented in this paper, and are considered in the following paragraphs.

(i) *Are most repeat sequences internally similar in structure to satellite DNA sequences?*

Figure 2 of this paper provides a diagrammatic description of the sea urchin repeats that is directly inconsistent with the character of several known satellite DNAs. Several of these, e.g. guinea pig satellite (Southern, 1970), mouse satellite (Sutton & McCallum, 1972; Biro *et al.*, 1975), certain *Drosophila* satellites (Gall & Atherton, 1974; Endow *et al.*, 1975; Brutlag & Peacock, 1975; Cordiero-Stone & Lee, 1976) have been shown to consist of lengthy tandem arrays of very short repeating sequence elements, even though (as in the case of mouse satellite (Southern, 1975)) a longer tandem periodicity may also be evident. In contrast, most of the length of the sea urchin repeat sequences investigated here is devoid of any internal repetitions, and no simple "ancestral" sequence elements can be convincingly discerned, though they may once have been present. This statement can be made for repeat sequences that have multiplied recently in evolution, as well as for very ancient repeat sequences such as that carried in CSp2108 (see Scheller *et al.*, 1981; Moore *et al.*, 1978,1981; Anderson *et al.*, 1981 for discussion of the evolution of these sequences). This result implies that satellite DNA and middle repetitive DNA are fundamentally different.

(ii) *Are most of the repetitive sequences in the genome divergent variants of one or a small number of canonical sequences?*

This proposition requires that statistically significant sequence homology exists among the divergent repeat sequence families, including those that do not react with each other under given conditions of renaturation. The sequences presented here are inconsistent with this requirement. No statistically significant homologies are shared by all or most of the sequenced repeats. Nor, for example, are any of them homologous to any extent with sequences belonging to the "Alu" repeat family of human DNA (Jelinek *et al.*, 1980). The most significant intersequence homology observed, that between CSp2108 sequence and the CSp2137 sequence (Fig. 4), is too weak to promote stable duplex formation in standard solution reactions. It follows that the eight repeat sequences studied here belong to eight distinct repeat families. There are no reasons to assume that the sequenced clones are exceptional in their lack of interrelationship. Therefore this argument indicates (as pointed out in Results) that the renaturation kinetics of repetitive sea urchin DNA must directly reflect the complexity of the repeat families, rather than reflecting the presence of a large number of variously divergent sequences related to a small number of canonical types.

(iii) *Do most repeat sequences belong to the class of "foldback" DNA?*

This proposition is *prima facie* unlikely since the amount of "foldback" DNA is usually much less than the amount of repetitive sequence. In the sea urchin genome, for example, the fraction of S_1 nuclease-resistant or hydroxyapatite-bound DNA isolated after very low C_0t incubation of fragments sheared to a few hundred nucleotides is only about 3%, while the amount of repetitive DNA is about 25% (Graham *et al.*, 1974). Similar ratios were obtained for *Xenopus* DNA (Davidson *et al.*, 1973) and human DNA (Schmid & Deininger, 1975). The sequences of the eight repeats presented here show that none of them could form stable foldback structures as defined by hydroxyapatite binding or S_1 nuclease resistance of sheared DNA at extremely low C_0t. However, inverse repeats that are separated by up to a few thousand nucleotides of other sequence are not uncommon (Wilson & Thomas, 1974: Schmid *et al.*, 1975) and these can only be observed after intrastrand renaturation of long DNA fragments.

We now consider the proposition that the stems of structures formed from such separate inverse repeats were included among the cloned duplex fragments whose sequences are presented in Figure 1. In sea urchin DNA the amount of S_1 nuclease-resistant intrastrand foldback formed from fragments several kb long is only 3 to 5% (unpublished data), again a small fraction of the total amount of repetitive sequence in the genome. Furthermore, the detailed studies of Anderson *et al.* (1981) show that in the 2108 and 2109 repeat families multiple family members do not usually occur in proximity to each other, in either inverse or direct orientation. This is not true of CSp2034 sequences, where both orientations have been observed for these elements within several kb of DNA (Anderson *et al.*, 1981; and unpublished data). Thus, though the cloned CSp2034 sequence could indeed have been a "separated inverse repeat" stem, this is unlikely in general for the cloned repeats, and in particular for CSp2108 and CSp2109. In sum, it seems probable that most sea urchin repeat elements are not internally palindromic to an extent that can be recognized under the experimental conditions usually applied, nor are they generally to be included in the intrastrand duplex fraction formed from separated inverted sequence elements.

(iv) *Can most repetitive sequences randomly isolated from the genome be accounted for as regions of known classes of repetitive genes?*

It is clear that both the complexity and the amount of the repeats in the genome are far too high for a large fraction of them to be accounted for as ribosomal or transfer RNA gene sequences (including the spacers associated with these species). The number of gene families coding for small nuclear RNAs (i.e. the number of snRNA species) is as yet not well determined, though it has been reported that there are 30 to 50 prominent snRNAs in sea urchin (Nijhawan & Marzluff, 1979) and human genomes (Zieve & Penman, 1976; Lerner *et al.*, 1980). This again suggests a low complexity compared to that of the total repetitive sequence class. Furthermore, repeat sequences in the sea urchin genome are mainly represented in large nuclear RNAs rather than small nuclear RNAs (Costantini *et al.*, 1978; Scheller *et al.*, 1978). However, a similar argument is more difficult to make

regarding the possibility that genomic repeat sequences are simply portions of multiple copy structural genes, since the number of these that exist is unknown. Three prior observations have made this possibility less attractive. First, the size of many repeat families (e.g. see Klein *et al.*, 1978) seems too large for them to be plausibly regarded as sets of related structural genes. Secondly, a great number of the interspersed repeat sequences in the genomes of many animals are only a few hundred nucleotides long. Thirdly, both complements of structural gene sequences are not expected to be represented in cellular RNAs, in contrast to all the interspersed sea urchin repeat sequences studied (Costantini *et al.*, 1978; Scheller *et al.*, 1978). The sequence data presented here provide specific evidence that at least some repeat sequences are not translatable. As discussed in Results (see Table 3), most of the sequenced repeats contain multiple translation arrest signals in every reading frame of both strands. It is of course possible that other members of these repeat families are permissive for translation since the various members are not identical in sequence. For the CSp2034 and CSp2112 sequences, however, this would require differences in the sequence resulting in the abolition of at least six and three particular stop codons, respectively, with no additional termination signal added in that reading frame. This is not likely since Klein *et al.* (1978) showed that the amount of divergence within both these families is low. The frequency of sequence change (alterations per nucleotide) needed to open a reading frame in the CSp2112 sequence, for example, would be more than five times the average frequency of sequence differences observed between 2112 family members (4%).

Though the repeat sequences are probably not included in codogenic message regions, they may be included in many mRNAs as non-translated regions and they are very likely included in a large number of nuclear mRNA precursors. Costantini *et al.* (1980) showed that repeat transcripts are present on over 65% of the poly(A) RNA molecules of sea urchin eggs. These molecules are the immediate precursors of, or are synonymous with, the maternal message loaded on polysomes after fertilization. Most nuclear RNA molecules are known to contain interspersed repeats in sea urchins (Smith *et al.*, 1974; Scheller *et al.*, 1978) as in other organisms (Darnell & Balint, 1971; Holmes & Bonner, 1974; Federoff *et al.*, 1977; see review in Davidson & Britten, 1979). The repeat transcripts occur in sea urchin eggs and in embryo and adult nuclear RNAs in a stage and tissue-specific manner. We regard the solution of the problem of understanding the role of these transcripts as a key to understanding the physiological significance of genomic repetitive sequences.

(b) *Significant internal sequence features*

Statistically significant though very short direct and inverse repeat features were found in five of the eight sequences studied. This may be a common characteristic of repetitive sequences. The meaning of this observation is unknown. The fact that these features are statistically significant of course does not necessarily imply that they are functional in any sense (though they could be), and since they are not present in all repeats they are clearly not a "necessary" attribute of repetitive sequences. "Statistical significance" merely indicates that the occurrence of these features in sequences of the given lengths by means of random assortment of

nucleotides is extremely improbable. Therefore the internal short repeat features probably indicate some aspect of DNA evolution which results in duplication or nearby insertion of very short blocks of pre-existing sequence.

(c) *Alternative sequence arrangements within the repeat family*

Figure 6 of this paper demonstrates the existence of sequence subelements of the 2109 family that are about 20 to 40 nucleotides in length. These subelements are defined by their presence in several genomic family members, where they are terminated by different flanking sequence subelements. Figure 7 shows that a somewhat larger scale subelement insertion (or deletion) has also occurred in a 2109 family member. Here a sequence a few hundred nucleotides long is present in one family member and absent in another. Another example is found in the recent studies of genomic sequences related to the CSp2108 sequence carried out by Scheller *et al.* (1981). This investigation revealed a large assemblage of about 1000 partly homologous sequences constituting a repeat "superfamily". Each 2108 sequence is several kb in length, and is composed of many sequence subelements that occur in a given order in given subfamilies, and in different orders in other subfamilies. These sequence subelements could be up to several hundred nucleotides in length. We conclude that repeat sequence subelements ranging in size from a few nucleotides up to a few hundred nucleotides occur in various arrangements among related repeats in the sea urchin genome. This is probably a common occurrence, since it was observed in two of the three families we have investigated in this respect, *viz.* in the 2109 and 2108 families. However, alternative subelement arrangements are probably absent in the 2034 sequence family, the various sequences of which appear mainly colinear (Anderson *et al.*, 1981; and unpublished data).

Alternative sequence subelement arrangement is a newly discovered property of repeat families or superfamilies. Wensink *et al.* (1979) recently described a similar property in some *Drosophilia* genomic repeats which like the "scrambled" 2108 repeats considered by Scheller *et al.* (1981) are relatively long sequences. Similarly, Jones *et al.* (1979) have observed alternative sequence subelement arrangements in several families of silkmoth chorion genes. The observations presented here show that alternative sequence subelement order occurs in short interspersed repeats as well. The calculation given in Table 5 demonstrates that the alternative subelement order partly explains a prominent physical property of renatured 2109 repeats, *viz.* their low thermal stability. It remains to be seen to what extent random sequence divergence and to what extent alternative sequence subelement order accounts for the observed decrease in thermal stability generally observed in repeat sequence heteroduplexes. Both phenomena clearly are involved. The significant consequence is that in cases such as the 2109 family the amount of actual sequence divergence between homologous subelements of different family members is much less than calculated on the basis that divergence alone accounts for their reduced heteroduplex stability. For example, block I of λ2109B-18 contains 42 nucleotides which are only 2·4% different from the corresponding region of CSp2109. In other words, short sequence elements may be as well

conserved within interspersed repeat families as are longer sequence elements in families displaying high heteroduplex thermal stability.

An interesting speculative aspect of the alternative arrangement shown in Figure 6 is that except for the final 16 nucleotide segment of block 2 the subelement arrangement in λ2109B-18 is a circular permutation of that in CSp2109. It is known that sea urchin repeat families possess some relatively rapid means of dispersion to distant regions of the genome during evolution (Moore *et al.*, 1978,1981; Anderson *et al.*, 1981). One speculative interpretation of this fact is that repeat sequence evolution involves free circular DNA elements capable of insertion in various genomic regions. It may be relevant that Stanfield & Lengyel (1979) have reported that *Drosophila* cells contain closed circular episomal DNA enriched in genomic repetitive sequences.

Whatever the evolutionary mechanism by which alternative sequence subelement order occurs, small repeat sequence subelements could be the real units of repeat sequence evolution and perhaps function. The subelements in the examples we describe here are relatively conserved in sequence while the whole interspersed repeat unit is not. If repeat sequences (or their transcripts) participate in macromolecular interactions, the many different possible combinations and permutations of subelements could provide an enormously expanded functional diversity.

We thank Dr Norman Davidson for his helpful and critical review of this manuscript, and Dr James Bonner for the use of his computer. We acknowledge the technical assistance of Mr Robert Gimlich. This research was supported by National Institutes of Health grant GM-20927. Two authors (J.W.P. and R.H.S.) were supported by a National Institutes of Health National Research Service award (GM-07616), and one author (D.M.A.) was supported by a National Institutes of Health postdoctoral fellowship (HD-05510).

REFERENCES

Anderson, D. M., Scheller, R. H., Posakony, J. W., McAllister, L. B., Trabert, S. G., Beall, C., Britten, R. J. & Davidson, E. H. (1981). *J. Mol. Biol.* **145**, 5–28.

Biro, P. A., Carr-Brown, A., Southern, E. M. & Walker, P. M. B. (1975). *J. Mol. Biol.* **94**, 71–86.

Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. & Chambon, P. (1978). *Proc. Nat. Acad. Sci., U.S.A.* **75**, 4853–4857.

Britten, R. J., Graham, D. E. & Neufeld, B. R. (1974). In *Methods in Enzymology.* (Grossman, L. & Moldave, K., eds), vol. 29E. pp. 363–406, Academic Press, New York.

Britten, R. J., Cetta, A. & Davidson, E. H. (1978). *Cell*, **15**, 1175–1186.

Brutlag, D. L. & Peacock, W. J. (1975). In *The Eukaryote Chromosome* (Peacock, W. J. & Brock, R. D., eds), pp. 35–45, Australian National University Press, Canberra.

Cordiero-Stone, M. & Lee, C. S. (1976). *J. Mol. Biol.* **104**, 1–24.

Costantini, F. D., Scheller, R. H., Britten, R. J. & Davidson, E. H. (1978). *Cell*, **15**, 173–187.

Costantini, F. D., Britten, R. J. & Davidson, E. H. (1980). *Nature (London)*, **287**, 111–117.

Darnell, J. E. & Balint, R. (1971). *J. Cell Physiol.* **76**, 349–356.

Davidson, E. H. & Britten, R. J. (1979). *Science*, **204**, 1052–1059.

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

Dykes, G., Bambara, R., Marians, K. & Wu, R. (1975). *Nucl. Acids Res.* **2**, 327–345.

Endow, S. A., Polan, M. L. & Gall, J. G. (1975). *J. Mol. Biol.* **96**, 665–692.

Federoff, N., Wellauer, P. K. & Wall, R. (1977). *Cell*, **10**, 597–610.

Gall, J. G. & Atherton, D. D. (1974). *J. Mol. Biol.* **85**, 633–664.
Graham, D. E., Neufeld, B. R., Davidson, E. H. & Britten, R. J. (1974). *Cell*, **1**, 127–137.
Hardison, R. C., Butler, E. T., III, Lacy, E., Maniatis, T., Rosenthal, N. & Efstratiadis, A. (1979). *Cell*, **18**, 1285–1297.
Holmes, D. S. & Bonner, J. (1974). *Proc. Nat. Acad. Sci., U.S.A.* **71**, 1108–1112.
Jelinek, W. R., Toomey, T. P., Leinwand, L., Duncan, C. H., Biro, P A., Choudary, P. V., Weissman, S. M., Rubin, C. M., Houck, C. M., Deininger, P. L. & Schmid, C. W. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 1398–1402.
Jones, C. W., Rosenthal, N., Rodakis, G. C. & Kafatos, F. C. (1979). *Cell*, **18**, 1317–1332.
Klein, W. H., Thomas, T. L., Lai, C., Scheller, R. H., Britten, R. J. & Davidson, E. H. (1978). *Cell*, **14**, 889–900.
Laskey, R. A. & Mills, A. D. (1977). *FEBS Letters*, **82**, 314–316.
Lerner, M. R., Boyle, J. A., Mount, S. M., Wolin, S. L. & Steitz, J. A. (1980). *Nature (London)*, **283**, 220–224.
Mandel, M. & Marmur, J. (1968). In *Methods in Enzymology* (Grossman, L. & Moldave, K., eds), vol. 12B, pp. 195–206, Academic Press, New York.
Maxam, A. M. & Gilbert, W. (1977). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 560–564.
Moore, G. P., Scheller, R. H., Davidson, E. H. & Britten, R. J. (1978). *Cell*, **15**, 649–660.
Moore, G. P., Costantini, F. D., Posakony, J. W., Davidson, E. H. & Britten, R. J. (1980). *Science*, **208**, 1046–1048.
Moore, G. P., Pearson, W. R., Davidson, E. H. & Britten, R. J. (1981). *Chromosoma*, in the press.
Needleman, S. B. & Wunsch, C. D. (1970). *J. Mol. Biol.* **48**, 443–453.
Nijhawan, P. & Marzluff, W. F. (1979). *Biochemistry*, **18**, 1353–1360.
Ohmori, H., Tomizawa, J. & Maxam, A. M. (1978). *Nucl. Acids Res.* **5**, 1479–1485.
Sanger, F. & Coulson, A. R. (1978). *FEBS Letters*, **87**, 107–110.
Schaffner, W., Kunz, G., Daetwyler, H., Telford, J., Smith, H. O. & Birnstiel, M. L. (1978). *Cell*, **14**, 655–671.
Scheller, R. H., Thomas, T. L., Lee, A. S., Klein, W. H., Niles W. D., Britten, R. J. & Davidson, E. H. (1977*a*). *Science*, **196**, 197–200.
Scheller, R. H., Dickerson, R. E., Boyer, H. W., Riggs, A. D. & Itakura, K. (1977*b*). *Science*, **196**, 177–180.
Scheller, R. H., Costantini, F. D., Kozlowski, M. R., Britten, R. J. & Davidson, E. H. (1978). *Cell*, **15**, 189–203.
Scheller, R. H., Anderson, D. M., Posakony, J. W., McAllister, L. B., Britten, R. J. & Davidson, E. H. (1981). *J. Mol. Biol.* **149**, 15–39.
Schmid, C. W. & Deininger, P. L. (1975). *Cell*, **6**, 345–358.
Schmid, C. W., Manning, J. E. & Davidson, N. (1975). *Cell*, **5**, 159–172.
Smith, M. J., Hough, B. R., Chamberlin, M. E. & Davidson, E. H. (1974). *J. Mol. Biol.* **85**, 103–126.
Southern, E. M. (1970). *Nature (London)*, **227**, 794–798.
Southern, E. M. (1975). *J. Mol. Biol.* **94**, 51–69.
Stanfield, S. W. & Lengyel, J. A. (1979). *Proc. Nat. Acad. Sci., U.S.A.* **76**, 6142–6146.
Sutton, W. D. & McCallum, M. (1972). *J. Mol. Biol.* **71**, 633–656.
Swartz, M. N., Trautner, T. A. & Kornberg, A. (1962). *J. Biol. Chem.* **237**, 1961–1967.
Thiery, J.-P., Macaya, G. & Bernardi, G. (1976). *J. Mol. Biol.* **108**, 219–235.
Wensink, P. C., Tabata, S. & Pachl, C. (1979). *Cell*, **18**, 1231–1246.
Wetmur, J. G. (1976). *Ann. Rev. Biophys. Bioeng.* **5**, 337–361.
Wilson, D. A. & Thomas, C. A., Jr (1974). *J. Mol. Biol.* **84**, 115–144.
Zieve, G. & Penman, S. (1976). *Cell*, **8**, 19–31.

Chapter III

Reprint Series
30 May 1980, Volume 208, pp. 1046-1048

SCIENCE

Evolutionary Conservation of Repetitive Sequence Expression in Sea Urchin Egg RNA's

Gordon P. Moore, Franklin D. Costantini, James W. Posakony, Eric H. Davidson, and Roy J. Britten

Evolutionary Conservation of Repetitive Sequence Expression in Sea Urchin Egg RNA's

Abstract. *Cloned repetitive DNA sequences were used to determine the number of homologous RNA transcripts in the eggs of two sea urchin species,* Strongylocentrotus purpuratus *and* S. franciscanus. *The eggs of these species contain different amounts of RNA, and their genomes contain different numbers of copies of the cloned repeats. The specific pattern of repetitive sequence representation in the two egg RNA's is nonetheless quantitatively similar. The evolutionary conservation of this pattern suggests the functional importance of repeat sequence expression.*

The sea urchin genome contains 10^5 to 10^6 repetitive sequence elements belonging to several thousand nonhomologous families. These repeat families are represented in nuclear RNA's and in egg RNA in a manner specific to the state of differentiation (*1*, *2*). The individual repetitive sequence fragments were isolated by S1 nuclease digestion of partially renatured DNA from *Strongylocentrotus purpuratus*, and cloned by the addition of chemically synthesized restriction enzyme recognition sequences (*3*). The cloned fragments were labeled and strand-separated, and were used as probes to detect RNA's homologous to individual repeat families. The results can be summarized as follows: (i) All of nine cloned repeat sequences studied are represented in nuclear RNA's and egg RNA, and at least 80 percent of the various repeat families in the genome are represented in egg RNA. This is in marked contrast to single copy sequences, of which only a minor fraction are found in nuclear or egg RNA's. (ii) Each repeat sequence family is represented to a particular extent in each RNA. The sequence concentrations of transcripts complementary to particular cloned repeats may differ more than 100-fold in a given RNA, and different families are highly represented in each RNA investigated. Thus the levels of representation are a function of the state of differentiation. (iii) Both strands of each cloned repeat are represented in all the RNA's studied. The complementary repeat transcripts in general reside on different RNA molecules, and probably derive from asymmetric transcription of separate multiple genomic copies oriented oppositely (*4*). A number of repeat elements of each sequence family are probably utilized in transcription. Recent studies (*5*) have shown that many of the single copy maternal messenger RNA transcripts stored in the egg are linked covalently with short repetitive sequence elements, and most of the repeat transcripts are associated with this interspersed RNA fraction.

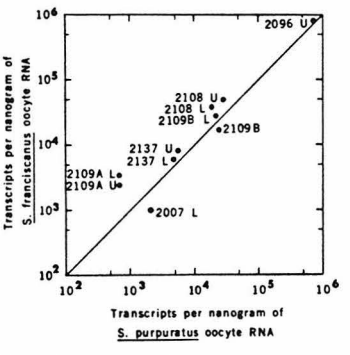


Fig. 1. Comparison of repeat sequence concentrations in the RNA of *Sf* and *Sp* eggs. Data are from *S. purpuratus* (abscissa) and *S. franciscanus* (ordinate) of Table 1. The solid line with slope 1 is the equivalence curve that would represent equal transcript concentrations in the eggs of the two species.

We now report experiments demonstrating that the specific pattern of repeat sequence representation in egg RNA has been quantitatively conserved since the divergence of two sea urchin species. The lineages leading to these species, *S. purpuratus* (*Sp*) and *S. franciscanus* (*Sf*), apparently separated 15 to 20 million years ago (*6*). They are of particular interest since their genomes contain clearly different numbers of copies of many repeat sequences (*7*). An issue addressed in the following experiments is whether the pattern of repeat sequence expression reflects these large evolutionary changes in repetitive sequence family size, or is independent of them.

In Table 1 are listed the results of measurements made with six cloned repeat fragments. The two strands of each cloned repeat were separated and reacted individually with *Sp* and *Sf* egg RNA's. The fraction of the total egg RNA complementary to each probe sequence was calculated from either kinetic or titration measurements (legend to Table 1) [see (*1*), the source of much of the *Sp* data listed]. In order to estimate the number of specific RNA transcripts, the mass of RNA per egg was multiplied by the specific fraction of the RNA hybridizing with the cloned probe. This calculation is of interest because the diameter of the *Sf* egg is about 120 μm compared to 80 μm for *Sp*, and thus the *Sf* egg is three times larger in volume. Measurements by the phloroglucinol method (*8*) show that the mass of RNA per average *Sf* egg is 8.3 ng compared to about 3 ng for the *Sp* egg (*9*). Thus the concentration of total RNA (most of which is ribosomal RNA) in the eggs of the two species is almost the same. In Table 1 the number of transcripts complementary to each cloned probe per egg is given for *Sp* and *Sf*, respectively; these values have been normalized for the egg RNA content of each species to give the number of specific transcripts per nanogram of total RNA. Measurement of transcript prevalence by kinetic or titration analysis is accurate only to within a factor of about 2.

The conclusions (*1*) for *Sp* egg RNA hold as well for *Sf* egg RNA (Table 1). Apart from clone CSp2096, which in several ways is atypical (*10*), the specific representation of the individual clones differs in *Sf* egg RNA by factors as large as 50. Both strands of all but one cloned repeat are also represented in *Sf* egg

0036-8075/80/0530-1046$00.50/0

RNA, just as in *Sp* egg RNA. The exception is CSp2096, the representation of which is highly asymmetric in the eggs of both species (*10*).

The major result is that the pattern of representation of the six cloned probes in egg RNA in essentially similar in the two species (Table 1). The data on transcripts per oocyte show that the sequences highly represented in *Sp* egg RNA are also highly represented in *Sf* egg RNA. The number of specific transcripts per unit of RNA mass were normalized for the egg RNA content of the two species. With one exception (CSp2109A) the number of transcripts per unit of RNA mass is the same within a factor of 2 in *Sp* and *Sf* eggs, while the size of the genomic repeat sequence families represented by this set of clones differs as much as ninefold. In the case of the CSp2109A sequence, the *Sf* egg has three to four times more homologous transcripts per nanogram of egg RNA, but the *Sp* genome has 20 times the number of homologous repeats as the *Sf* genome. When one DNA strand is more highly represented in RNA, this transcriptional asymmetry is retained in the eggs of both species. This is clear in the case of CSp2096, and is probably also evident for the CSp2108 sequence. The degree of proportionality in the numbers of specific transcripts per egg for the two

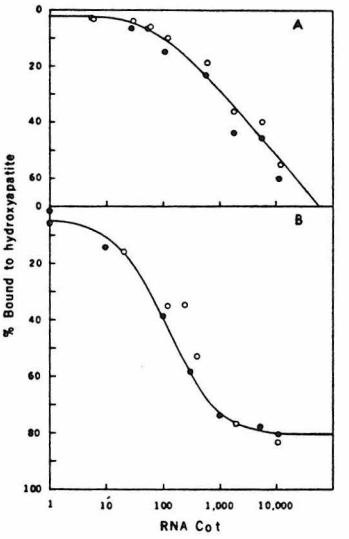


Fig. 2. Hybridization of *S. purpuratus* repetitive ^{3}H-labeled DNA with excess *S. franciscanus* and *S. purpuratus* egg RNA. (A) Hybridization of total *S. purpuratus* repetitive ^{3}H-labeled DNA tracer with *Sp* and *Sf* egg RNA. The *S. purpuratus* DNA was labeled in vivo to a specific activity of 1.1×10^6 cpm/μg. The ^{3}H-labeled DNA was denatured, reassociated to a C_0t of 40, treated with S1 nuclease, and chromatographed on Sepharose CL-2B. Included material of mode size 300 nucleotides was hybridized with a 10^6-fold mass excess of *S. purpuratus* egg RNA (●) and *S. franciscanus* egg RNA (○) at 55°C in 0.4*M* phosphate (Na^+) buffer. The fraction of ^{3}H-labeled DNA in RNA-DNA hybrids was assayed by binding to hydroxyapatite at 50°C in 0.12*M* phosphate (Na^+) buffer and elution at 99°C. The tracer was that described by Costantini *et al.* (*1*). The data were pooled and fit with the assumption of two kinetic components (solid line). The line shown is a least squares solution based on two kinetic components. This analysis suggests that 38.6 percent of the ^{3}H-labeled DNA repeat tracer hybridizes with a second-order rate constant of $1.79 \times 10^{-3} M^{-1}$ sec^{-1} and 38.5 percent hybridizes with a second-order rate constant of $5.64 \times 10^{-5} M^{-1}$ sec^{-1}. (B) Hybridization of highly expressed *S. purpuratus* repetitive ^{3}H-labeled DNA with *Sp* and *Sf* egg RNA. The *Sp* repetitive ^{3}H-labeled DNA tracer used in (A) was fractionated to select for those repeat sequences whose transcripts are most abundant by preparative reassociation with *Sp* egg RNA. This tracer was twice reassociated to RNA C_0t 500 with a 3×10^4-fold excess of *Sp* egg RNA as described (*1*). The ^{3}H-labeled DNA which bound to hydroxyapatite after the two RNA excess hybridizations constituted 19 percent of the repetitive tracer. The average sequence concentration of the RNA's represented in this tracer was 10^5 copies per egg. This selected ^{3}H-labeled DNA fraction was hybridized with a 10^6-fold excess of *Sp* (●) and *Sf* (○) egg RNA, and the kinetics of the reaction was determined. Hybridizations and assay conditions were as in (A). On the basis of a single second-order kinetic component, the second-order rate constants were $7.9 \times 10^3 M^{-1}$ sec^{-1} and $3.5 \times 10^3 M^{-1}$ sec^{-1}, respectively. The solid line shown is the least squares solution of the *Sp* hybridization points.

Table 1. Abundance of transcripts complementary to six cloned repetitive sequences in oocyte RNA's of *S. purpuratus* and *S. franciscanus*.

Clone	Strand*	Length† (NT)	Genomic‡ frequency *S. purpuratus*	Transcripts per oocyte§		Transcripts per nanogram of RNA		Genomic frequency ratio *S. purpuratus*/ *S. franciscanus*††
				S. purpuratus‖	*S. franciscanus*¶	*S. purpuratus***	*S. franciscanus***	
CSp2007	U	1,100	400	7,000	N.D.	2,100	N.D.	9
	L			7,000	8,300	2,100	1,000	
CSp2108	U	204	20‡‡	97,000	420,000	29,000	50,000	0.7
	L			59,000	310,000	18,000	37,000	
CSp2109A	U	180	900	2,500	20,000	800	2,400	20
	L			2,800	29,000	800	3,500	
CSp2109B	U	111	200	83,000	140,000	25,000	17,000	4
	L			73,000	240,000	22,000	29,000	
CSp2137	U	226	530	19,000	68,000	5,800	8,200	3
	L			16,000	50,000	4,800	6,000	
CSp2096	U	203	80	2,400,000	6,000,000	720,000	720,000	2.5
	L			Low§§	Low§§	Low§§	Low§§	

*The two complementary strands of each cloned repeat were designated "upper" (U) and "lower" (L) according to their electrophoretic mobility on neutral polyacrylamide gels (*2*). †The length of each cloned repetitive element except CSp2007 is known from DNA sequence analysis (*11*). The length of CSp2007 was estimated from electrophoretic mobility of the duplex fragment (*12*). ‡Haploid genomic repetition frequency was estimated from the kinetics of reassociation of each cloned repeat with an excess of *S. purpuratus* DNA (*12*). §The limit of detection of RNA transcripts is dependent on the specific activity of the cloned repeat DNA probe. The ^{32}P-labeled DNA probes used in this study were routinely labeled to about 10^7 cpm/μg, which allows detection of about 100 transcripts per nanogram of egg RNA. ‖The number of RNA transcripts complementary to a cloned DNA sequence was determined by comparing the rate of hybridization of labeled, strand-separated cloned repeats to the rate of hybridization of single copy ^{3}H-labeled DNA with excess oocyte RNA (*13*). The kinetics of the RNA-driven hybridizations with the cloned repeats are approximately second order since both complementary strands are present in the RNA. Titration measurements were carried out by reacting increasing amounts of oocyte RNA with a constant quantity of labeled, strand-separated DNA tracer. Data analysis was as described (*1*, *2*). The values presented under *S. purpuratus* are averages of previous determinations (*1*) and new determinations. The latter are one determination for the CSp2108 upper strand, one determination for the CSp2109B lower strand, and the determination shown for the CSp2096 upper strand. These three measurements were obtained by the kinetic method. The rate constants (M^{-1} sec^{-1}) were 2.5 × 10^{-3} for the CSp2108 upper strand, 4.5 × 10^{-3} for the CSp2109B lower strand, and 2.2 × 10^{-1} for the CSp2096 upper strand. ¶The values shown were determined by the kinetic method except in the case of CSp2109A upper strand and one determination of CSp2137 upper strand, which were obtained by the titration method. The second-order rate constants (M^{-1} sec^{-1}) were: CSp2108 U, 8.0 × 10^{-3}; CSp2108 L, 5.9 × 10^{-3}; CSp2109A U, 5.3 × 10^{-4}; CSp2109B U, 4.0 × 10^{-3}; CSp2109B L, 7.0 × 10^{-3}; CSp2137 U, 1.1 × 10^{-3}; CSp2137 L, 9.8 × 10^{-4}; CSp2007 L, 3.4 × 10^{-4}; CSp2096 U, 2.2 × 10^{-1}. In the titration measurements the fractions of the total RNA found to consist of homologous transcripts were for CSp2109A, U, 3.1 × 10^{-5}; and for CSp2137 U, 1.1 × 10^{-6}. **The number of RNA transcripts per nanogram of oocyte RNA complementary to a cloned DNA sequence is the number of transcripts per egg divided by 3.3 (nanograms of RNA per *Sp* egg) or 8.3 (nanograms of RNA per *Sf* egg). ††Data from (*6*, *14*). ‡‡Recent work (*15*) has shown that there are many sequences distantly related to CSp2108 in the genome of *Sp*. The value shown refers to the size of a discrete subset of closely related sequences, which also exist in *Sf*. §§The case of CSp2096 L is complicated by the asymmetry of the RNA transcripts. During hybridization, labeled DNA probe must compete for its scarce RNA complement with the more abundant RNA strand. We estimate that the maximum prevalence of the CSp2096 L transcript is 7 percent of the CSp2096 U prevalence, or about 50,000 transcripts per nanogram of egg RNA in both species.

species is shown in Fig. 1, where the quantitative similarity in the pattern of repeat sequence expression in the two egg RNA's extends over the full range of transcript sequence concentrations.

The results in Table 1 and Fig. 1 derive from only six cloned repeat sequences and could be nonrepresentative. We attempted to generalize these results by using repetitive DNA tracers prepared from whole sheared *Sp* DNA. As a control, in Fig. 2A the reactions of such a tracer with *Sp* and *Sf* egg RNA's are shown. The kinetics of these reactions are almost indistinguishable, and it follows that the overall distribution of repeat transcript sequence concentrations in the two egg RNA's is similar. The generalization that those repeat families highly represented in *Sp* egg RNA are also highly represented in *Sf* egg RNA is tested in Fig. 2B. A subfraction of the genomic repeat tracer enriched for the sequences prevalent in *Sp* egg RNA was reacted with both egg RNA's. The experiment shows that the repeat sequences prevalent in *Sp* egg RNA are represented in about the same concentrations in *Sf* egg RNA. We conclude that the quantitative similarity in the patterns of repeat sequence representation is general in *Sp* and *Sf* egg RNA's.

The distribution of repeat sequence concentrations in egg RNA is almost certainly physiologically significant. This distribution is highly sequence specific, and different distributions have been found in other RNA's that have been studied. Specific processes of repetitive sequence transcription and RNA accumulation must operate during oogenesis. Our results show that these processes have been adjusted during evolution so as to preserve the repeat transcript sequence concentrations despite significant changes in genomic repeat family sizes and in egg volume and total RNA content. Conceivably all the repeat sequences belonging to expressed families are transcribed in both genomes, but the turnover and transcription rates have altered so as to compensate for the changes in repeat family size. The following argument suggests an alternative view. The particular pattern of repeat sequence representation present in the eggs of both of these species probably existed in their common ancestor as well. Therefore only a subset of the repeats of some repetitive families may be transcribed during oogenesis, since in the cases where the *Sp* families are larger than the *Sf* families this is not reflected in higher *Sp* transcript sequence concentrations. Possibly only repetitive sequence elements of given families which are present in the genomes of both species are expressed. The general implication would be that although multiple homologous repeats are utilized in transcription, these may often represent a subset of the whole genomic sequence family.

Our results emphasize the functional importance of repetitive sequence transcription. The particular pattern of repeat sequence representation in mature eggs is quantitatively conserved between two sea urchin congeners, although the size of the various repeat families in their genomes differs significantly. The repeat sequences themselves have diverged less than single copy sequences during the evolution of this genus (*7*). Conservation of repeat sequence expression suggests that the mechanism restraining change in the primary sequences is evolutionary selection, based on important but unknown functions of the repetitive RNA transcripts.

GORDON P. MOORE
FRANKLIN D. COSTANTINI
JAMES W. POSAKONY
ERIC H. DAVIDSON
Division of Biology, California Institute of Technology, Pasadena 91125
ROY J. BRITTEN
Kerckhoff Marine Laboratory, California Institute of Techology, Corona del Mar 92625

References and Notes

1. F. D. Costantini, R. H. Scheller, R. J. Britten, E. H. Davidson, *Cell* **15**, 173 (1978).
2. R. H. Scheller, F. D. Costantini, M. R. Kozlowski, R. J. Britten, E. H. Davidson, *ibid.*, p. 189.
3. R. H. Scheller *et al.*, *Science* **196**, 197 (1977).
4. This statement is based on the observation that the repeat transcripts of both nuclear and egg RNA's are interspersed with single copy regions in the RNA [M. J. Smith, B. R. Hough, M. E. Chamberlin, E. H. Davidson, *J. Mol. Biol.* **85**, 103 (1974); (*5*)]. However, single copy sequences are always asymmetrically represented in sea urchin RNA's according to studies with both cloned and total DNA tracers [Z. Lev *et al.*, *Dev. Biol.*, in press; B. R. Hough, M. J. Smith, R. J. Britten, E. H. Davidson, *Cell* **5**, 291 (1975); (*5*)]. Therefore, to account for the presence of complementary copies of each repeat sequence, it is necessary to postulate the transcription of multiple repeat elements. The hybrid duplexes formed when cloned repeats are reacted with RNA exhibit heterogeneous thermal stabilities, suggesting that transcription occurs from at least several members of the repeat family (R. H. Scheller, D. M. Anderson, R. J. Britten, E. H. Davidson, unpublished observations).
5. F. D. Costantini, R. J. Britten, E. H. Davidson, in preparation.
6. J. W. Durham, in *Treatise on Invertebrate Paleontology(U) Echinodermata*, R. C. Moore, Ed. (Geological Society of America and Univ. of Kansas Press, New York, 1966), vol. 3, part 1, p. 270.
7. G. P. Moore, R. H. Scheller, E. H. Davidson, R. J. Britten, *Cell* **15**, 649 (1978).
8. Z. Dische and E. Borenfreund, *Biochim. Biophys. Acta* **23**, 639 (1957).
9. A. H. Whiteley, *Am. Nat.* **83**, 249 (1949).
10. The CSp2096 sequence could be a fragment of some highly expressed structural gene or could be related in sequence to such a gene. However, the representation of both complements, and the short length of most of the repeat sequence elements found in egg RNA, argue strongly against the possibility that the repeat transcripts in general code for proteins (*1*). Furthermore, six out of eight repetitive sequence fragments that have been sequenced (*11*) display translation termination signals in all three reading frames of both strands. CSp2096 is exceptional in this regard. In this fragment two reading frames of the strand represented in RNA lack termination codons, though the other strand is translationally blocked in all three frames.
11. J. W. Posakony, R. J. Britten, E. H. Davidson, in preparation.
12. W. H. Klein *et al.*, *Cell* **14**, 889 (1978).
13. B. R. Hough-Evans, B. J. Wold, S. G. Ernst, R. J. Britten, E. H. Davidson, *Dev. Biol.* **60**, 258 (1977).
14. G. P. Moore, W. R. Pearson, E. H. Davidson, R. J. Britten, in preparation.
15. D. M. Anderson *et al.*, in preparation.
16. Supported by NIH grant HD-05753, NIH biomedical research support grant RR-07003, NIH postdoctoral training grant GM-07401, and NIH postdoctoral fellowship GM-07290 (to G.P.M.) and NIH predoctoral training grant GM-07616 (to J.W.P.).

26 December 1979; revised 19 February 1980

Chapter IV

Structural and Developmental Characteristics of Interspersed Polyadenylated RNA's of Sea Urchin Eggs and Embryos

INTRODUCTION

A very large fraction of the sea urchin genome consists of repetitive sequence elements interspersed with single-copy DNA (Graham et al., 1974). These repeats number over 10^5 and belong to several thousand more or less distinct sequence families (Graham et al., 1974; Klein et al., 1978; this thesis, Chapter II). Members of individual interspersed repeat families are found widely scattered in the genome (this thesis, Chapter I).

The expression of interspersed repetitive sequences in sea urchin RNA has recently begun to be studied in detail. Hybridization experiments (Costantini et al., 1978; Scheller et al., 1978; this thesis, Chapter III) have demonstrated the presence in egg total RNA and in gastrula stage and adult intestine nuclear RNA's of transcripts complementary to a set of cloned repetitive sequence elements (each representing a different repetitive sequence family). Different repeat sequences were found to be represented at different levels in each RNA, and the representation of a given repeat sequence in the three RNA's showed a characteristic variation. Moreover, it was observed that both complements of each repeat are present in the RNA, generally at approximately equal prevalence levels. The size of the RNA molecules bearing these repetitive sequence transcripts was estimated by hybridization of cloned repeat probes to egg total RNA and gastrula nuclear RNA fractionated on denaturing DMSO-sucrose gradients (Costantini et al., 1978; Scheller et al., 1978). It was found that most of the RNA's complementary to the four probes used were 1000-2500 nucleotides (nt) long; however, evidence of RNA degradation in these experiments suggested that this might be an underestimate. Nevertheless, it was clear that the repeat-bearing RNA's are at least several-fold longer than the cloned repeat elements themselves.

The nature of repetitive sequence transcripts in the polyadenylated fraction of sea urchin egg RNA was investigated by Costantini et al. (1980). Electron microscope visualization of annealed egg poly(A)$^+$ RNA revealed multi-stranded branched

structures involving approximately 65% of the RNA mass. These structures appeared to consist of short regions of RNA duplex separated by longer single-stranded regions. Based on the earlier finding (Costantini et al., 1978) that both complements of several cloned repeat elements are represented in egg total RNA, the duplex regions could be interpreted as renatured repetitive sequences. That complex networks involving several RNA molecules were formed by annealing suggested that many poly(A)$^+$ RNA species in the egg carry more than one repeat sequence. Another experiment measured the kinetics of hybridization of a fraction of egg poly(A)$^+$ RNA enriched in repeat-containing transcripts with excess whole sea urchin DNA. The analysis indicated that 85-90% of the mass of the enriched fraction consisted of single-copy sequence transcript. It was concluded that between 35% and 70% of egg poly(A)$^+$ RNA has an interspersed sequence organization, in which short repetitive sequences are covalently linked to longer single-copy sequences.

The present studies were undertaken to investigate in detail the structural and developmental characteristics of interspersed polyadenylated transcripts in sea urchin eggs and embryos. Experiments described below demonstrate that the two complements of a given repeat sequence are represented on different sets of poly(A)$^+$ RNAs, and that these transcripts are comparatively large; i.e., greater than 3 kilobases (kb) in length. The polyadenylated RNAs of three embryonic stages of sea urchin development are shown to contain multiple, large transcripts complementary to each of several cloned repeat sequences, and evidence of clear quantitative and probably qualitative developmental variation in repeat transcripts is presented. The interspersed nature of individual polyadenylated transcripts bearing specific repeat sequences is demonstrated by analysis of cDNA clones selected with two different repeat probes. The cloned regions are shown to consist of a repetitive segment complementary to the probe repeat, flanked by single-copy segments or by segments containing other, non-homologous repeat sequences. Single-copy fragments

from three of the cDNA clones are found to react with single poly(A)$^+$ transcripts in egg RNA, and these transcripts are also found in the embryo, though in two cases an additional transcript appears at a specific developmental stage. Finally, it is shown that there is little similarity in size or apparent prevalence of individual repeat-containing transcripts in the poly(A)$^+$ egg RNA of two closely related sea urchin species, S. purpuratus and S. franciscanus.

MATERIALS AND METHODS

(a) Growth of sea urchin embryos

Gametes of Strongylocentrotus purpuratus and S. franciscanus were collected following intracoelomic injection of 0.5 M KCl. Eggs were washed by repeated settling in Millipore-filtered seawater (MPFSW).

Embryos of S. purpuratus were grown at 15-16°C at a density of 0.5-1x10^4/ml in MPFSW containing either 30 U/ml penicillin plus 50 µg/ml streptomycin or 40 µg/ml gentamicin. The embryo suspension was constantly stirred and was aerated with bubbles of compressed air. Development was initiated by fertilizing eggs at the stated density with 10 ml/liter of a 1% suspension of fresh sperm in MPFSW. Only cultures showing >98% fertilization and morphologically normal development were used for RNA preparations.

(b) Preparation of egg and embryo poly(A)$^+$ RNA

Total poly(A)$^+$ RNA of eggs and embryos was prepared as follows. All solutions were treated by vigorous shaking with 0.02% diethylpyrocarbonate (DEP) followed by autoclaving. All glassware was baked overnight in a 180°C oven, and all plastic-ware was rinsed with 0.02% DEP in glass-distilled water (dH_2O) and then autoclaved. Eggs or embryos in MPFSW were pelleted by centrifugation and washed by resuspension and pelleting through ice-cold calcium-magnesium-free seawater (Hynes and Gross,

1970) brought to pH 3 with citric acid. The pellets were suspended in a buffer (Hough-Evans et al., 1977) containing 7 M urea, 50 mM sodium acetate (pH 5.1), 10 mM EDTA, 15 mM EGTA, 1% SDS, and 10 µg/ml polyvinyl sulfate, and homogenized with 5-10 strokes of a "B" pestle in a Dounce homogenizer. RNA was purified from this homogenate by the cesium chloride pelleting method of Glisin et al. (1974). The homogenate was diluted 1:1 with DEP-treated dH_2O, and solid CsCl (1 g/ml) was added and dissolved. The homogenate was transferred to Beckman Quick-Seal tubes and underlaid with a cushion of 5.7 M CsCl in 10 mM sodium acetate (pH 5.1). Centrifugation was for 24 h at 39 K rpm, 20°C, in a 70 Ti or 60 Ti rotor. The RNA pellets were dissolved in DEP-treated dH_2O at 4°C and precipitated twice with two volumes of ethanol at -20°C from 0.3 M sodium acetate pH 5.5. The RNA was again dissolved in dH_2O and precipitated overnight at -20°C with two volumes of 4.5 M sodium acetate (pH 5.5) (Childs et al., 1979). The sodium acetate precipitation was effective in removing the considerable amounts of non-ultraviolet-absorbing carbohydrate contaminating the RNA of later embryos. The final pellet was dissolved in dH_2O and stored at -70°C.

Poly(A)$^+$ RNA was isolated from total RNA by two cycles of binding to oligo(dT)-cellulose (Aviv and Leder, 1972) at room temperature, using the buffers suggested by the manufacturer (Collaborative Research). The RNA was dissolved in dH_2O and stored at -70°C.

A note on intactness of poly(A)$^+$ RNA preparations: it was occasionally observed that a preparation of poly(A)$^+$ RNA examined by standard methods [e.g., electrophoresis in agarose gels containing methyl mercury (Bailey and Davidson, 1976)] and appearing largely intact was nonetheless unusable for the analysis of large (>5 kb) repeat-containing transcripts by virtue of being slightly degraded. This became evident upon carrying out gel blot hybridizations of the RNA with cloned repeat elements as probes: the discrete bands usually observed (particularly those representing the largest transcripts)

were obviously diminished in intensity, the relative intensity of lower molecular weight bands was emphasized, and the apparent "background" was higher.

Obviously, the appearance on RNA gel blots of large, rare transcripts is affected most by slight, presumably random, degradation of this kind; the appearance of small, prevalent transcripts is affected least. Thus, if two RNA populations seem to differ in the presence or absence of a given transcript, it should be shown that the RNA appearing to lack the transcript nonetheless contains intact larger (and preferably rarer) transcripts than that under consideration. The point may seem trivial, but the literature contains increasingly frequent "demonstrations" of RNA intactness in which ribosomal RNA (!) or other short, prevalent transcripts are used as standards of integrity for longer, rarer species (e.g., Dawid et al., 1981; Zuker and Lodish, 1981).

(c) Preparation and screening of a cDNA library to egg poly(A)$^+$ RNA

A library of cDNA clones was prepared by C. Flytzanis from S. purpuratus egg poly(A)$^+$ RNA purified by three passages over oligo(dT)-cellulose. The construction of the library was essentially according to Rowekamp and Firtel (1980), except that synthesis of the first strand of the cDNA by reverse transcriptase was randomly primed by oligonucleotides of calf thymus DNA (Taylor et al., 1976). Following alkali treatment to remove the RNA and purification of the cDNA on Sephadex G-100, the second strand was synthesized using DNA polymerase I. The resulting double-stranded cDNA was treated with S-1 nuclease and size-selected by chromatography on Biogel A-150M. The selected double-stranded cDNA was tailed with poly(dC) and annealed with the plasmid vector pBR322 tailed at the PstI site with poly(dG). The annealed DNA was then used to transform E. coli strain HB 101 by the method of Cohen et al. (1972).

Storage and screening of bacterial colonies containing the recombinant cDNA plasmids was carried out essentially as described by Hanahan and Meselson (1980).

To select cDNA clones bearing specific repeat sequences, the library was screened with the cloned repetitive sequence elements CS2109A and CS2137 (Klein et al., 1978). Hybridization was carried out in DNA hybridization buffer (see below) at 50° or 60°C. The hybridized filters were washed at 60°C in 5X SET, 0.1% SDS [1X SET is 0.15 M NaCl, 0.03 M Tris-HCl (pH 8.0), 2 mM EDTA].

(d) Preparation of plasmid and genomic DNAs

Bacteria containing recombinant plasmids were grown and the plasmids amplified according to the method of Norgard et al. (1979), and plasmid DNAs were then isolated as described previously (Scheller et al., 1977).

High-molecular weight DNA from sea urchin sperm or testes was prepared by W. Pearson and J. Roberts following the procedure of Britten et al. (1974).

(e) Restriction enzyme digestion and labeling of DNA

DNA was digested with various restriction enzymes using a simplified buffer system, as follows:

H-M-L-K Buffer System

H buffer	M buffer
0.15 M NaCl	50 mM NaCl
6 mM Tris-HCl pH 8.0	6 mM Tris-HCl pH 8.0
6 mM $MgCl_2$	6 mM $MgCl_2$
L buffer	**K buffer**
6 mM NaCl	10 mM KCl
6 mM Tris-HCl pH 8.0	10 mM Tris-HCl pH 8.0
6 mM $MgCl_2$	10 mM $MgCl_2$

Special buffers:

Eco RI buffer	Taq I buffer
50 mM NaCl	6 mM NaCl
0.1 M Tris-HCl pH 7.4	6 mM Tris-HCl pH 8.5
6 mM $MgCl_2$	6 mM $MgCl_2$

The buffers were kept as 10X stocks at 4°C. Reducing agents were added to the digestion mixtures as needed. All digestions were carried out at 37°C, with the exception of Taq I digestions, which were at 65°C under mineral oil.

DNA of each of the cDNA clones studied here was treated with a set of 24 restriction enzymes; digestions were carried out in the buffers shown below:

Enzyme	Buffer
Ava I	M
Ava II	M
Bam HI	H
Bcl I	M
Bgl I	M
Bgl II	M
BstE II	M
Cla I	L
Dde I	H
Eco RI	Eco RI
Hinc II	M
Hind III	M
Hinf I	M
Kpn I	L
Nae I	M

Pst I	M
Pvu I	H
Pvu II	M
Rsa I	M
Sal I	H
Sau 3AI	M
Taq I	Taq I
Xba I	H
Xho I	H

Bam HI, Bgl II, Eco RI, Hinc II, Hind III, Pst I, and Sal I were prepared by Maria Alonso; the remainder were from New England Biolabs. The cDNA clone inserts were mapped in situ in the plasmid by a combination of single and double restriction enzyme digests.

Restriction enzymes leaving 5' overhanging bases were chosen preferentially, to permit the use of "filling in" (repair synthesis) by Klenow polymerase as a labeling method. In initial experiments, DNA was dissolved in L buffer (above) containing 5 mM dithiothreitol (DTT) and whichever α-^{32}P-dXTP's (Amersham; aqueous, 2000-3000 Ci/mmol) were required to repair the ends created by the restriction enzyme(s). The labeled triphosphates were present in sufficient quantity to repair all the ends. Following addition of 1-2 units Klenow polymerase (Boehringer), the mixture was incubated at room temperature for 20 min, after which all four cold dXTP's were added to 40 µM (each) and incubation was continued for 10 additional minutes. The reaction was terminated by the addition of EDTA to 20 mM. In later experiments, labeling was routinely performed by adding DTT, labeled dXTP's, and Klenow polymerase directly to restriction enzyme digest mixtures in M buffer. The final buffer was 1/2X M plus 5 mM DTT. The remaining steps were as above.

(f) Isolation of labeled DNA restriction fragments

Labeled DNA fragments were separated on 4% or 6% polyacrylamide slab gels in Tris-borate-EDTA buffer (Maxam and Gilbert, 1977) run at 200-250 V (regulated) in a 4°C cold room. For isolation of duplex fragments, Klenow polymerase labeling mixtures (see above) were loaded directly on the gel. Labeled repetitive elements to be strand-separated were denatured by incubation in a boiling water bath for 5 min and quenched in an ice/water bath before loading.

Following electrophoresis, regions of the gel containing the desired DNA fragments were localized by autoradiography and excised. The DNA was eluted by incubating crushed gel pieces in 2 ml 0.1 M NaCl, 10 mM Tris-HCl (pH 7.4), 1 mM EDTA overnight at 37°C. Acrylamide was removed by centrifuging the elution mixture through silanized glass wool in a 5 ml plastic syringe to which a plastic tube had been attached to receive the liquid containing the DNA. The labeled fragments were stored at -20°C.

(g) Agarose gel electrophoresis and nitrocellulose blotting of DNA and RNA

Restriction enzyme digests of genomic DNA were electrophoresed (5-7 µg/lane) on 1% agarose gels in Tris-acetate-EDTA buffer. Hind III-digested λ DNA and Hinf I-digested pBR322 DNA (Sutcliffe, 1979) were used as size markers. Digests of clone DNA were run on 1.5%-2.5% agarose gels, using pBR322/Hinf I and vector fragments as markers. Following ethidium bromide (1 µg/ml) staining and gel photography, the DNA was transferred to nitrocellulose paper by the method of Southern (1975), using 10X SET as the transfer medium. After blotting (8-12 h) the filter was washed briefly in 3X SET, air dried, and baked for 2-3 h at 80°C in a vacuum oven.

Poly(A)$^+$ RNA (1-2 µg/lane) was electrophoresed on 0.8% agarose gels containing 20 mM sodium morpholinopropane sulfonic acid (MOPS; pH 7.0), 5 mM sodium acetate, 1 mM EDTA, and 2.2 M formaldehyde (Rave et al., 1979; Lehrach et al., 1977). RNA samples were denatured in 60% formamide, 2.2 M formaldehyde in the above

MOPS-acetate-EDTA buffer for 5 min at 65°C. The running buffer was MOPS-acetate-EDTA and was constantly recirculated with a peristaltic pump during electrophoresis. RNA size markers, run in parallel lanes, were as follows: E. coli 16S (1541 nt; Brosius et al., 1978) and 23S (2904 nt; Brosius et al., 1980) rRNAs, sea urchin 18S (∿1840 nt; F. Costantini, unpublished) and 26S (∿3850 nt; F. Costantini, unpublished) rRNAs, and poliovirus virion RNA (∿7500 nt including poly(A); Kitamura et al., 1981). Poliovirus RNA was kindly provided by G. Larsen and E. Wimmer.

Following electrophoresis, lanes containing markers were sliced from the gel, stained 1-2 h in 1 μg/ml ethidium bromide in dH_2O, destained overnight at 4°C in dH_2O, and photographed under UV illumination. Transfer of RNA from the remainder of the gel to nitrocellulose filters (B. Seed, personal communication) was accomplished as follows. The gel was equilibrated for 20-40 min in one gel volume 20X SSC (1X SSC = 0.15 M NaCl, 0.015 M sodium citrate, pH 7.5) with occasional agitation. Using 20X SSC as the transfer medium, the RNA was blotted onto nitrocellulose paper with the same blotting set-up as for DNA (Southern, 1975). After blotting (10-12 h), the filter was soaked in 20X SSC, air dried, and baked for 2-3 h at 80°C in a vacuum oven.

(h) DNA and RNA gel blot hybridization

DNA gel blots were preincubated in DNA hybridization buffer [5X SET, 1X Denhardt's (Denhardt, 1966), 20 mM sodium phosphate buffer (PB; pH 6.8) and 50 μg/ml sheared, denatured E. coli or calf thymus DNA] for 2-4 h in sealed plastic bags on a shaking water bath at the hybridization temperature (see below). Heat-denatured radioactive probe DNA was then added, and hybridization was carried out for 14-20 h. When the fragment to be used as a probe contained a poly (dG)·(dC) region from one end of a cDNA clone (see above), poly(rC) at 15 μg/ml was present during both the preincubation and hybridization. The hybridized filters were washed 2-3X in 5X SET, 0.1% SDS at 60°C.

Blot hybridizations with genomic DNAs (genome blots) were carried out at 60°C. Hybridizations with clone DNA on the filter (clone blots) were carried out at either 50° or 60°C.

RNA gel blots were pre-treated in 10X Denhardt's, 5X SET, 50 mM PB for 2-3 h in sealed plastic bags on a shaking water bath at the hybridization temperature (42°C). The pretreatment buffer was then removed and replaced with RNA hybridization buffer [5X SET, 1X Denhardt's, 20 mM PB, 50 µg/ml sheared, denatured E. coli or calf thymus DNA, and 20-50% formamide (see below)], and the filters were pre-incubated for 2-3 h at 42°C. As above, poly(rC) at 15 µg/ml was included if necessary. Heat-denatured radioactive probe DNA was added, and hybridization was carried out for 14-18 h at 42°C. The hybridized filters were washed 2-3X at 60°C in 5X SET, 0.1% SDS.

The reaction criterion of RNA gel blot hybridizations in different experiments was varied by varying the concentration of formamide in the hybridization buffer (Casey and Davidson, 1977). For hybridization of sequences well-matched to the probe sequence (such as when single-copy fragments were used as probes), 50% formamide was used. To permit efficient hybridization of divergent sequences (e.g., poorly matched repeats), the concentration of formamide was lowered to 20%. Interspecies transcript comparisons (Figure 8) were carried out in 40% formamide.

(i) Electron microscopy of annealed RNA

Egg poly(A)$^+$ RNA was denatured, annealed, and spread for electron microscopy from 80% formamide as described by Costantini et al. (1980). Single-stranded ØX174 DNA was included as an internal size standard (Glass and Wertz, 1980).

(j) DNA sequencing

DNA sequencing was carried out by the method of Maxam and Gilbert (1980), using duplex restriction fragments asymmetrically labeled by repair synthesis with Klenow polymerase (see above).

RESULTS

(a) RNA gel blot patterns of repeat-containing transcripts in S. purpuratus egg poly(A)$^+$ RNA

A principal goal of these studies was to examine the sizes and relative prevalences of polyadenylated transcripts deriving from individual repetitive sequence families, and to compare these parameters for the two complements of each repeat.

In Figure 1 are displayed autoradiograms of gel blots of S. purpuratus egg poly(A)$^+$ RNA, hybridized at reduced criterion (see Materials and Methods) with the separated strands of cloned repetitive sequence elements representing four different interspersed repetitive sequence families (see Klein et al., 1978; this thesis, Chapter II; see also Chapter I; Scheller et al., 1981a). A number of observations can be made: (1) The repetitive sequence probes are complementary to multiple discrete transcripts of various sizes. (2) The patterns of bands revealed by hybridization to the two separated strands of a given repeat are different, demonstrating that the two complements of that repeat sequence are carried on different molecules. (3) The transcripts complementary to each repeat are relatively large; nearly all are greater than 3 kb in length, with a rough number average size of perhaps 5-6 kb. Thus, virtually all of the transcripts detected are more than ten times as long as the cloned repeat sequence used as a probe. (4) The much greater intensity of some bands in the autoradiograms compared to others indicates that certain transcript species bearing a given repeat sequence are much more prevalent in the poly(A)$^+$ egg RNA population than are other species bearing the same repeat (though this conclusion must be drawn cautiously; see below). That only a small number of intensely hybridizing bands is observed implies that only a few members of a given repeat family are transcribed into prevalent polyadenylated RNA's. Conversely, the apparent light "background" of hybridization in these experiments is at least in part indicative of a large number of individual RNA species which are comparatively rare.

The repetitive sequence probes used in these experiments represent repeat families which exhibit a considerable degree of intrafamilial sequence divergence, particularly the 2109A and 2109B families (see Klein et al., 1978; Scheller et al., 1981a). The population of transcripts deriving from such a repeat family could thus be expected to include sequences both well and poorly matched to any individual cloned member of the family. Consequently, to permit detection of the more distantly related sequences, the blot hybridizations were carried out at a reaction criterion significantly reduced from that routinely used to detect sequences well-matched to the probe sequence (see Materials and Methods). Even under the reduced criterion conditions, however, the extent of hybridization to the various transcripts—and hence the relative intensity of the various bands in the autoradiogram—will still depend somewhat on which cloned representative of the repeat family is used as a probe. To test the magnitude of this effect, blot hybridizations were carried out with the separated strands of two different cloned DNA's representing the 2109B repeat family; the two probes are substantially divergent from each other (see legend to Figure 1). In Figure 1, the resultant autoradiograms are compared. Quantitative differences in the intensity of particular bands are evident, but the overall patterns are qualitatively very similar. Thus it seems likely that the detection of the diverse set of transcripts bearing 2109B repeat family sequences has been reasonably complete. Moreover, the interpretation of the large variations in band intensity as being indicative of differences in the prevalences of specific transcripts is well supported.

(b) Repeat-containing poly(A)$^+$ RNA's in embryonic stages of sea urchin development

The developmental pattern of expression of individual repetitive sequence families was studied by hybridizing gel blots of total poly(A)$^+$ RNA of the egg and three embryonic stages—16-cell, hatched blastula, and mid-gastrula—with the separated strands of four different cloned repetitive sequence elements. The autoradiograms are displayed in Figure 2.

It will be seen from these experiments that the general characteristics of repeat-containing poly(A)$^+$ transcripts in embryos are very similar to those described above for egg repeat-containing RNAs. Thus, the polyadenylated RNA of all three embryonic stages contains multiple discrete transcripts of various sizes complementary to each repeat probe; different sets of transcripts carry the two complements of each repeat; and the detectable transcripts are almost all greater than 3 kb in length.

RNA from 16-cell embryos was included in these experiments because at this stage the very great majority of even the non-ribosomal RNA mass consists of transcripts inherited from the egg, with little contribution from new synthesis in the embryo (Ernst et al., 1980). Nevertheless, four cleavages have occurred and five hours have elapsed since fertilization, and the opportunity thus exists to examine the fate, with respect to size and relative prevalence, of maternally inherited repeat-containing transcripts. Figure 2 shows that the pattern of gel bands hybridizing to the various repeat sequence probes in 16-cell poly(A)$^+$ RNA is qualitatively similar to that in egg poly(A)$^+$ RNA, although some differences are detectable. Thus, although these experiments do not provide a quantitative assessment of the fate of particular transcripts, it is clear that there is no large-scale degradation and/or processing of maternal repeat-containing transcripts in early cleavage, at least up to 5 h post-fertilization.

By the hatched blastula or midgastrula stages of development (approximately 18 h and 40 h, respectively, after fertilization), very substantial differences are evident in the pattern of poly(A)$^+$ RNAs hybridizing with most of the repeat probes, compared to egg poly(A)$^+$ RNA (Figure 2). Some of these differences are due to the developmental disappearance or decline in prevalence of a transcript species prominent in egg and 16-cell RNAs. Turnover and/or processing of particular maternal transcripts would account for these changes. A much more common type of difference between the repeat-containing RNA populations of the egg and of later embryos is the appearance of qualitatively new transcript species or an increase in the prevalence

of species already present in the egg. Most of these changes are clearly the result of embryonic transcription, especially given the fact that at the blastula and gastrula stages of sea urchin development, a very large fraction of the non-ribosomal RNA mass has been synthesized since fertilization (Davidson, 1976). However, the relatively modest developmental increase in the apparent prevalence of some species could possibly be due to post-fertilization addition of poly(A) to maternal transcripts (Wilt, 1977), resulting in their more quantitative isolation on oligo(dT)-cellulose. Finally, many of the gel bands detectable in egg RNA with the various repeat probes appear to persist at approximately the same intensity in the blastula and gastrula patterns. These may represent maternal transcripts which survive intact in the later embryo, or species whose embryonic transcription more or less balances the turnover of maternal molecules.

In summary, these experiments show that as early sea urchin development proceeds, the set of transcripts representing an individual repetitive sequence family may participate in the same kinds of qualitative and quantitative modulation as observed for various mRNAs; e.g., those for the histones (Childs et al., 1979), the actins (Crain et al., 1981), and the tubulins (Alexandraki and Ruderman, 1981).

(c) Isolation and mapping of repeat-containing cDNA clones of egg poly(A)$^+$ RNA

To permit the detailed study of specific repeat-bearing transcripts, a cDNA clone library of egg poly(A)$^+$ RNA was constructed (see Materials and Methods). Based on the results presented in the previous sections, which indicated a large size (>3 kb) for the repeat-containing transcripts in egg and embryo polyadenylated RNA's, the cDNA destined for cloning was synthesized by "random priming" with oligonucleotides from calf thymus DNA, in order to provide a more uniform representation of the sequences in such long transcripts than would have been obtained by oligo(dT) priming at the 3'-terminal poly(A) tail. The double-stranded cDNA was inserted

into the PstI site of pBR322 by the poly(dG)·poly(dC) homopolymeric extension method (Rowekamp and Firtel, 1980).

Sets of cDNA clones bearing specific repeats were selected from the library by screening with two different cloned repetitive sequence elements, CS2109A and CS2137 (referred to below as "CS repeats"). These cloned elements and the genomic repeat families they represent have been characterized extensively in previous studies (Klein et al.,1978; this thesis, Chapters I and II; Scheller et al., 1981a; Moore et al., 1981). The 2109A family has at least 900 members in S. purpuratus and is organized in the genome as individual short repetitive sequence elements flanked on either side by single copy sequences (this thesis, Chapter I). Thus, it is a typical short interspersed repeat family. Less is known about the organization of the 2137 family, but the available evidence indicates that it consists of short (Moore et al., 1981) repetitive elements which are scattered in the genome (see below). The 2137 family has about 530 members in the S. purpuratus genome (Klein et al., 1978).

Four cDNA clones selected with each of the two CS repeats were subjected to detailed restriction enzyme mapping (Figure 3). Restriction fragments bearing sequences homologous to the CS repeat were identified by gel blot hybridizations of the restriction digests used for mapping. The minimal region of each clone to which the CS repeat sequence could be localized is indicated in Figure 3 by an asterisk. In most cases the region is short compared to the total length of the cloned cDNA insert, and similar in size to the CS repeat itself (of course, the actual homology may be even shorter than shown in the figure; see below). In some of the clones, the CS repeat region is located at one end of the cDNA insert, but in others (e.g., p2109A-2) it is flanked on both sides by non-CS sequences.

(d) Sequence analysis of a transcribed repetitive sequence

The CS repeat region of p2109A-2 was examined in detail by determining its

nucleotide sequence. As shown in Figure 4, the homology to CS2109A is short (62 nt) but well-matched (94%). It is possible that additional homologous sequences have been removed from this region by RNA processing (or conceivably by artifactual recombination in the cDNA clone), but it is more likely that the particular member of the 2109A repeat family represented in p2109A-2 is simply a truncated version. Chapter II of this thesis demonstrates that members of the 2109A and B repeat families are not simply colinear variants of each other, but instead may differ by the presence or absence of various sequence elements. It is noteworthy that the sequence enclosed in brackets in Figure 4 is followed immediately in CS2109A by one element of a large inverse repeat (beginning at nucleotide 69 of CS2109A; see Figure 3 of Chapter II) which is not present in p2109A-2.

The symmetrical representation of repetitive sequences in sea urchin RNA's led to the suggestion (e.g., Scheller et al., 1978) that repeat sequence transcripts do not play a role in the protein coding function of messenger RNA. More specific evidence on this point was presented in Chapter II, where it was shown that most repeat sequences may be untranslatable due to the presence of translation termination codons in all three phases of both strands. However, as pointed out there, it remained possible that those members of a repeat family which are actually transcribed might contain open translational reading frames. The sequence shown in Figure 4 represents clear evidence that this is not true in the specific case of p2109A-2. The region (nucleotides 88 to 234) which has been sequenced in both directions is translationally blocked in all six possible reading frames; additional termination codons are found in the remainder of the sequence. Thus, that portion of the p2109A-2 transcript represented in Figure 4 is non-translatable.

(e) Sequence organization of cloned regions of repeat-containing egg transcripts

The sequence organization of the cDNA clones was studied using isolated,

radioactively labeled restriction fragments from each clone as hybridization probes. The various fragments are indicated in Figure 3 below the restriction map of the source clone; each fragment is given a designation denoting the number of the source clone and the restriction enzyme(s) used to generate it. Thus, 10HB is a fragment from p2109A-10 delimited by a Hind III site and a Bgl II site.

The probe fragments were each reacted with two different kinds of gel blots: restriction enzyme digests of total genomic DNA of two S. purpuratus individuals and restriction digests of the cDNA clones in which the cDNA insert was carried on a single fragment (see legend to Figure 5). The genome blot experiments allowed us to determine whether a given cDNA fragment contains repetitive sequences or only single-copy sequence. The clone blot experiments showed whether the fragment contains sequences shared by the other clones or only sequences unique to the source clone. The results of these blot hybridization experiments are presented in Figure 5 (A, B, D, and E) and summarized in Figure 3 and Table I. For comparision purposes, genome blots probed with the two CS repeats are also shown (Figure 5A and B).

In all four of the p2109A cDNA clones, restriction fragments flanking the CS repeat region each react with at most two different-sized genomic fragments (Figure 5A). For reasons discussed below, we interpret the two bands in the DNA of a given individual as two different alleles rather than two nonallelic copies of the sequence. Thus, these cDNA clone fragments consist entirely of single-copy sequence (denoted in Figure 3A by a thin line representing the fragment). In the case of p2109A-2, it was possible to show that sequences on both sides of the CS repeat region (fragments 2A and 2HH) are single-copy in the genome. By contrast, restriction fragments overlapping or encompassing the CS repeat region (e.g., 2AH) react with the entire spectrum of fragment sizes in the genome blots, just as does the CS2109A repeat itself. (These repeat-containing fragments are denoted in Figure 3A by a thick line representing the fragment.)

It is important to note that the single-copy restriction fragments from different p2109A cDNA clones react with different fragments of genomic DNA, implying that these sequences are not shared between clones. This was verified directly by the clone blot experiments, which showed that the single-copy fragments react only with their respective source clone (Figure 5D; results summarized in Table I). As expected, restriction fragments containing all or part of the CS repeat region react with all four clones (though most strongly with the source clone).

The organization of the p2137 cDNA clones is somewhat more complicated. As does the CS2137 repeat itself, restriction fragments from all four clones which overlap or contain the CS repeat region react with genomic DNA fragments of all sizes (Figure 5B). In three cases (p2137-5, -7, and -9), restriction fragments which flank the CS repeat region clearly contain other repetitive sequences (thick lines representing the fragments in Figure 3B). Only in CS2137-1 is a restriction fragment flanking the CS repeat region found to be single-copy in the genome (thin line representing the fragment in Figure 3B). The clone blot experiments (Figure 5E) show that the repetitive sequences flanking the CS repeat regions of p2137-5, -7, and -9 are not shared between clones, and that the single copy flanking sequence of p2137-1 is likewise unique to that clone. As expected, restriction fragments from the CS repeat regions themselves react with all four clones (results summarized in Table I).

The results of the experiments presented above may be summarized as follows: (1) The eight cDNA clones represent egg RNA transcripts in which a short repetitive sequence (a member of either the 2109A or the 2137 repeat family) is covalently linked to either single-copy sequence or other repetitive sequences (Figure 3). This is shown by the results of the genome blot experiments with selected restriction fragments from each clone (Figure 5A and B). (2) The four cDNA clones selected with each CS repeat represent different regions of the genome containing different

members of the repeat family. This is demonstrated by three lines of evidence: the dissimilarity of the restriction maps of the clones (Figure 3); the lack of homology between clones in the regions flanking the CS repeat region, as well as the greater hybridization signal always observed in the reaction of a CS repeat region fragment with its source clone compared to the other clones (Figure 5D and E); and finally, the observation that single copy sequences from different clones (p2109A-2, -10, -16, and -18, and p2137-1) react with different fragments of genomic DNA (Figure 5A and B).

(f) Polymorphism in genomic regions flanking transcribed repeats

The genome blot experiments shown in Figure 5A and B suggested that the regions of the genome which contain the single-copy sequences of interspersed transcripts may be quite polymorphic, inasmuch as single-copy fragments from the cDNA clones (Figure 3) generally reacted with different-sized restriction fragments in the two individual DNA's examined. This phenomenon was studied in more detail by carrying out genome blot experiments with the DNA's of 6 different S. purpuratus individuals, digested with two different restriction enzymes. The blots were hybridized with single-copy restriction fragments from five of the cDNA clones (p2109A-2, -10, -16, and -18, and p2137-1; see Figure 3). Representative autoradiograms are shown in Figure 6A-D; the overall results of these experiments are summarized in Table II.

Two points are immediately evident from Figure 6A-D. First, there is a high degree of interindividual polymorphism in the sizes of the Eco RI and Hind III restriction fragments of genomic DNA which contain the single-copy probe sequences. However, some common bands are observed, as in Figure 6A. In this case, the probe fragment (10 HB from p2109A-10; see Figure 3A) is bounded on one side by a site for the enzyme used in the genome digest (Hind III). In spite of this, at least eight different band sizes are evident in the six individuals (twelve haploid genomes).

Results similar to those shown in Figure 6A-D were obtained with the other three single-copy probes (Table II).

The second point to be noted is that each probe sequence reacts with two different-sized restriction fragments in the DNA of nearly every individual tested (Table II). The two bands in a given individual DNA are best interpreted as two different alleles of a sequence occurring once per haploid genome (i.e., the individual is heterozygous), since it is extremely unlikely that an individual would be homozygous for two non-allelic copies of the probe sequence, given the large amount of inter-individual polymorphism noted above.

The great variety of alleles present in a population of sea urchins can be visualized by genome blot experiments using mixtures of equal amounts of DNA from many individuals. Figure 6E shows the results of such an experiment, in which fragment 10 HB from p2109A-10 was used as a probe. The most frequent alleles are revealed by their intense hybridization signal, while less common variants are fainter. It can be seen that the pattern of alleles in two different populations is detectably different.

The possible significance of these polymorphisms will be considered in the Discussion.

(g) Individual interspersed transcripts in sea urchin development

The identification of single-copy restriction fragments within several of the repeat-containing cDNA clones (Figures 3 and 5) permitted an examination of the size and developmental characteristics of the specific interspersed transcripts represented by these clones. Gel blots of total poly(A)$^+$ RNA of the egg and of 16-cell, hatched blastula, and mid-gastrula stage embryos, hybridized with single-copy fragments from p2109A-10, -16, and -18, are shown in the first three panels of Figure 7. In egg poly(A)$^+$ RNA, each of the probe fragments reacts with a single transcript (presumably the template on which the corresponding cDNA was synthesized before cloning).

The transcript sizes are given in Table III. It is evident that the three cDNA clones from which the probe fragments were taken represent three different interspersed polyadenylated RNAs, each bearing a sequence belonging to the 2109A repeat family. Moreover, the sizes of these RNAs are within the range observed when either strand of the CS2109A repeat itself is hybridized to egg poly(A)$^+$ RNA (see Figure 1), suggesting that they are typical of transcripts homologous to this repeat sequence.

All three single-copy probe fragments reveal in all three embryonic RNAs a polyadenylated transcript of the same size as that found in egg poly(A)$^+$ RNA (Figure 7). The hybridizations to 16-cell stage RNA demonstrate in each case the persistence of the inherited maternal transcript through at least 5 h of embryonic development, though in the case of p2109A10 and -18, its prevalence appears to have declined somewhat. This may reflect the turnover of these transcripts or perhaps some processing event; in the latter case, however, the single-copy probe sequence is apparently not retained in the processed transcript, since no additional RNA band is detected in the blots.

Two of the single-copy probe sequences, those from p2109A-10 and -18, detect in the poly(A)$^+$ RNA of later embryonic stages a new transcript in addition to the predominant egg-type one: fragment 10 HB reveals a larger transcript at the mid-gastrula stage, fragment 18AR1 a smaller one in the hatched blastula (see Table III for sizes). Whether these additional RNA species have any kinetic relationship to the predominant transcript in each case (i.e., as a result of processing) or are instead novel primary transcripts from the same genomic regions is unknown.

The prevalence of all three of the maternal-type transcripts shown in Figure 7 is apparently significantly greater in blastula and/or gastrula stage embryos than in the egg. This result almost certainly indicates that the embryo genome is transcribed to produce qualitatively the same transcript in these three cases as is synthesized during oogenesis for storage in the egg. In support of this conclusion, two other

points should be mentioned here, with regard to the p2109A-10 transcripts. First, the appearance in the gastrula stage of a transcript larger than the maternal species clearly implies that new transcription from this locus has taken place in the embryo. Second, the maternal-type p2109A-10 transcript is easily detectable in total poly(A)$^+$ RNA prepared from pluteus stage embryos 84 h after fertilization (Figure 7). It is unlikely that inherited egg transcripts could persist for this long.

The last panel in Figure 7 shows a fourth RNA gel blot; the probe in this experiment was fragment 5A2 from p2137-5. Although this fragment carries sequences which are repetitive in the genome (see Figure 5B), hybridization to the gel blot under high stringency conditions (see Materials and Methods) reveals a single, comparatively rare transcript in egg poly(A)$^+$ RNA which is apparently very much more prevalent in hatched blastula poly(A)$^+$ RNA but not convincingly detectable in 16-cell and mid-gastrula stage RNAs. This again seems to be a repeat-containing transcript whose prevalence is developmentally modulated and which is synthesized in both oocytes and early embryos.

(h) Asymmetric representation of single-copy sequences in interspersed polyadenylated transcripts

Results presented in Figure 8B show that, in contrast to all of the repetitive sequences which have been studied (Costantini et al., 1980; see Figure 1), the single-copy sequences contained in the cDNA clones are represented asymmetrically in egg poly(A)$^+$ RNA. Two single-copy fragments from the same cDNA clone (results are shown for p2109A-16 and -18) were labeled on opposite strands and reacted with gel blots of egg poly(A)$^+$ RNA. In each case, only one of the two fragments was observed to hybridize significantly (see lanes marked with a "P"). This result also establishes the 5' -to- 3' polarity of the cDNA sequences in RNA (see Figure 3A).

(i) Evolutionary non-conservation of specific interspersed transcripts in egg poly(A)$^+$ RNA

The quantitative measurements presented in Chapter III of this thesis show that the overall concentration of transcripts complementary to several cloned repetitive sequence elements is very similar in the egg RNAs of two sea urchin species, S. purpuratus and S. franciscanus. This is in spite of the generally smaller size of the studied repeat sequence families in the genome of S. franciscanus (Moore et al., 1978).

These results raised the question of whether individual transcripts bearing a given repeat sequence are likewise evolutionarily conserved in their size and relative prevalence. Consequently, experiments were carried out in which gel blots containing equal amounts of egg poly(A)$^+$ RNA from the two species were hybridized at reduced criteria either with separated strands of cloned repeat sequences or with selected restriction fragments from the cDNA clones (Figure 3).

Figure 8A shows that the pattern of gel bands hybridizing to individual repeat sequences is markedly different in the two species. Thus, the discrete transcripts which are most prevalent in the egg polyadenylated RNA of one species are generally different in size from those which are most prevalent in the other species. In the case of CS2109A and the L strand of CS2137, there is a significant quantitative difference as well, in that the overall hybridization intensity to S. franciscanus egg poly(A)$^+$ RNA is much lower. (These results are not directly comparable with those of Chapter III, which refer to total egg RNA.) The differences observable in Figure 8A cannot be attributed to extreme sequence divergence in the 2109A, 2109B, and 2137 repeat families between the two genomes. Moore et al. (1981) showed that despite the lower reiteration frequency of these sequences in S. franciscanus, the interspecies divergence they exhibited was quite limited, leading to at most a 6°C difference in the average thermal stability of heteroduplexes between the CS repeat probes and genomic DNA. (In fact, heteroduplexes between S. franciscanus DNA and CS2109B

were actually more stable on the average than those formed with S. purpuratus DNA.) Moreover, the results shown in Figure 8A for the U strand of CS2137 reveal discrete polyadenylated transcripts containing this sequence in S. franciscanus RNA which are at least as prevalent as those in S. purpuratus. The large size (>5 kb) of most of these transcripts argues strongly that degradation of the S. franciscanus egg RNA is not responsible for the present observations. In a separate control experiment (not shown), a coding sequence probe from a cloned S. purpuratus actin gene (SpG17; Durica et al., 1980) was hybridized on a gel blot to the same preparations of egg poly(A)$^+$ RNA used for the experiments of Figure 8. The principal actin mRNA species was found to be the same size in both RNAs (2.2 Kb; Crain et al., 1981), and its apparent prevalence was in fact greater in S. franciscanus.

To test for the evolutionary conservation of individual interspersed polyadenylated transcripts, selected fragments from the cDNA clones (Figure 3) were hybridized at relaxed stringency (see Materials and Methods and the legend to Figure 8) to S. purpuratus and S. franciscanus egg poly(A)$^+$ RNAs. The results are shown in Figure 8B.

No transcript complementary to the single-copy fragment 16 PB (Figure 3A) was detected in S. franciscanus RNA. Hybridization with another single-copy fragment, 10 HB (Figure 3A), revealed in S. franciscanus an apparently quite rare transcript (visible in the original autoradiogram at the position shown by the arrow), somewhat larger in size (4.9 kb vs. 4.1 kb) than that in S. purpuratus. A third single-copy fragment, 18AR1 (Figure 3A), hybridized to a transcript of the same size in both species; its apparent prevalence is much lower in S. franciscanus. Finally, no transcript was detectable in S. franciscanus RNA, under the conditions used, with the repetitive fragment 5A2 (Figure 3B).

That the sequences used as probes in these experiments are in fact present and detectable in the genome of S. franciscanus was demonstrated by carrying out appropriate genome blot experiments. The relevant results are shown in Figure 5C.

As is the case in S. purpuratus (Figure 5A and B), the sequences contained in fragments 10 HB and 16 PB are single-copy in the genome of S. franciscanus, while at least part of fragment 5A2 consists of repetitive sequence.

In summary, the evidence presented in this section shows that while the overall concentration of transcripts representing particular repeat families is similar in the total egg RNA of S. purpuratus and S. franciscanus (Chapter III), the size and/or prevalence of individual polyadenylated transcripts bearing these repeat sequences is in general not the same in the two species. The specific examples displayed in Figure 8B reveal that some genomic regions (those in cDNA clone fragments 16 PB and 5A2) yield transcripts detectable in the egg poly(A)$^+$ RNA of S. purpuratus only, and that while other genomic regions are transcribed during oogenesis in both species, their polyadenylated products in S. franciscanus are of lower prevalence (fragment 18AR1) or of lower prevalence and different size (fragment 10 HB). Thus, the contribution of these specific sequences to the population of polyadenylated RNAs stored in the egg does not seem to be under strong evolutionary constraint.

DISCUSSION

(a) Structural and organizational properties of repeat-containing polyadenylated transcripts

The experiments described in this chapter provide a direct demonstration of several structural and organizational properties of the repeat-containing polyadenylated RNAs of sea urchin eggs and embryos:

(1) Individual repetitive sequence families are represented in RNA by multiple, discrete polyadenylated transcripts which are derived from different regions of the genome and carry the sequences of different members of the repeat family. However, in many cases a comparatively small number (2-10) of individual transcripts account for most of the overall prevalence of a given repetitive sequence. This

means that only a small fraction of the members of some repetitive sequence families are transcribed into prevalent polyadenylated RNAs.

(2) The two complements of each repetitive sequence are represented on different sets of polyadenylated transcripts.

(3) Repeat-containing polyadenylated RNAs are generally quite large, with an estimated number average length of perhaps 5-6 kb. In this respect they are different from most mRNAs, whose average size is perhaps closer to 2 kb. For example, Flytzanis et al. (1982) found that of six sea urchin embryo cDNA clones selected as representing developmentally regulated mRNAs, none hybridized on RNA gel blots to polyadenylated transcripts larger than 2.2 kb (C. Flytzanis, personal communication). Estimates of mRNA size in other systems (Lewin, 1980) are in this same range.

(4) Many, and probably most, repeat-containing polyadenylated RNAs consist of short repetitive sequence elements covalently linked with longer single-copy sequences or with other repetitive sequences (see also Costantini et al., 1980). The short repetitive elements do not in general occur at the extreme ends of the transcripts, but may be flanked on both sides by single-copy sequence. As noted in the Introduction, electron microscopic evidence suggests that many interspersed polyadenylated transcripts carry more than one repeat sequence (Costantini et al., 1980; see Figure 9).

(5) The single-copy sequences present on interspersed polyadenylated transcripts are represented asymmetrically in RNA. Taken together with (1) and (2) above, this implies that, as suggested earlier (Scheller et al., 1978; Costantini et al., 1980), the symmetric representation of most repetitive sequences in sea urchin RNAs is a result of asymmetric transcription of different genomic regions in which members of a given repeat family occur in either orientation with respect to the transcribed DNA strand.

The properties described above are summarized in Figure 9. The drawings show two polyadenylated transcripts, each containing two short repetitive sequence elements flanked by either longer single-copy sequences or other repeat sequences. The single-copy sequences in the two transcripts are different, but both RNAs carry one complement of repeat sequence R1, and could react under suitable conditions to form a short RNA duplex such as those apparent in the electron micrograph (Figure 9; Costantini et al., 1980).

This picture of the repeat-containing polyadenylated RNAs of sea urchin eggs and embryos should be compared to the findings of Firtel and co-workers (Kindle and Firtel, 1979; Kimmel and Firtel, 1979) regarding the structure of certain interspersed poly(A)$^+$ mRNAs in Dictyostelium. In the best studied case, a single short repetitive sequence (designated M4) occurs at the 5' terminus of a set of mRNAs containing diverse single-copy sequences. Moreover, the M4 repeat is represented asymmetrically in poly(A)$^+$ RNA. Thus, although many different members of the M4 repeat family are transcribed as part of polyadenylated RNAs, it is always with a fixed location and orientation in the transcript, in sharp contrast to the sea urchin repeats we have described (this chapter; Costantini et al., 1978; 1980; Scheller et al., 1978).

The total poly(A)$^+$ RNA of human lymphocytes has recently been shown (Crampton et al., 1981) to contain transcripts of several different repetitive sequences, and at least some of these appear to be linked to single-copy sequence transcript.

(b) Developmental properties of polyadenylated repeat-containing transcripts

The developmental characteristics of polyadenylated repeat-containing transcripts have been studied here by RNA gel blot hybridization, using as probes either cloned repeat elements or fragments of cDNA clones (Figure 3). Thus, we have been able to compare at different developmental stages the entire set of polyadenylated

RNAs deriving from an individual repeat family (Figure 2) and also look for the presence or absence of specific transcripts as development proceeds (Figure 7).

The experiments with 16-cell stage RNA demonstrate the persistence of most of the maternally inherited poly(A)$^+$ repeat transcripts for at least the first five hours of development. It would be of great interest to determine whether specific subsets of these transcripts are distributed to the different blastomeres, since it is at the 16-cell stage that the micromeres are formed—four small cells into which the determinants for the formation of the larval skeleton are segregated. More generally, the intracellular location of inherited repeat-containing transcripts in early embryos is also unknown.

Hybridization of cloned repeat probes to gel blots of poly(A)$^+$ RNA from blastula and gastrula stage embryos reveals clear—and in some cases dramatic—developmental changes in the set of transcripts bearing a specific repeat sequence. In a number of cases, qualitatively new transcripts, undoubtedly synthesized by the embryo, seem to appear (though these may instead represent sharp increases in the prevalence of transcripts that are comparatively rare in the egg). Nevertheless, many of the transcript species observed in later embryos clearly occur in the egg as well. The specific transcripts represented by three of the cDNA clones were found to be present throughout early development. The apparent increase in their prevalence between fertilization and the blastula or gastrula stages almost certainly indicates that the embryo nucleus actively transcribes some of the same repeat-containing RNA species as are synthesized by the oocyte nucleus for storage in the egg. Thus, it becomes evident that the developmental pattern of repetitive sequence transcription in sea urchins resembles in a qualitative way the pattern observed for mRNA sequences; i.e., new embryonic synthesis of many of the same sequences found in maternal RNA (Galau et al., 1977), accompanied by the developmental appearance of new species of transcript (Childs et al., 1979; Crain et al., 1981).

Both Kimmel and Firtel (personal communication) and Zuker and Lodish (1981) have described developmentally regulated sets of polyadenylated RNAs in Dictyostelium, the members of each set bearing a particular repeat sequence (see Davidson and Posakony, 1982, for review).

(c) Polymorphism in genomic regions flanking transcribed repeats

A high degree of genetic polymorphism was detected in the genomic regions surrounding the transcribed repeat sequences studied here, but its significance is uncertain. For example, it is unknown whether this polymorphism is due to base sequence differences in the regions cut by the restriction endonucleases, or to larger-scale differences such as insertions or deletions. More importantly, it cannot be determined with certainty at this point whether the polymorphic regions are actually part of the sequences included in the transcripts represented by the cDNA clones or instead lie entirely outside the transcript sequences. This latter possibility is unlikely, however.

A substantial amount of sequence polymorphism ($\sim$ 4%) is observed in the total single-copy DNA of S. purpuratus individuals (Britten et al., 1978). However, a restriction enzyme analysis of cloned sea urchin actin genes (Scheller et al., 1981b) indicated that allelic restriction site differences may be largely absent from extensive (several kilobase) regions including and surrounding the structural genes. Thus a possible interpretation of the high degree of polymorphism detected in the present experiments is that the sequences contained in the cDNA clones are part of non-coding regions lying some distance from any structural gene sequences.

(d) Significance of repeat-containing polyadenylated transcripts

The properties of sea urchin repetitive sequence transcripts which have so far been described permit some consideration of their significance, however tentative.

Although a large fraction (Costantini et al., 1980) are polyadenylated (and we have considered only these in this chapter), repeat-containing transcripts appear to have at least two structural characteristics which are in contrast to those of messenger RNAs. The first is their large size. The experiments presented here permit a rough estimate of 5-6 kb as the number average length of polyadenylated repeat transcripts in sea urchin eggs and embryos. Though the transcripts of only a few repeat families were examined, virtually all of the RNA species detected are longer than 3 kb. The average size of mRNAs in most cells, on the other hand, is generally thought to be significantly smaller, approximately 2 kb (reviewed by Lewin, 1980), and this conclusion is well supported by studies of specific mRNAs. The second point of contrast between these large repeat-containing RNAs and probably most mRNAs (Lewin, 1980) is the very presence of repetitive elements interspersed with other sequences in the transcript. Moreover, both complements of these elements are represented nearly equally in the RNA populations studied (Costantini et al., 1978; Scheller et al., 1978; this thesis, Chapter III); it is demonstrated here that these are on different sets of molecules. Polysomal poly(A)$^+$ RNAs in human cells bearing sequences of the highly reiterated "Alu" family have been described recently (Calabretta et al., 1981), but not enough is known of their properties to compare them with the transcripts considered here. As we have seen, both the asymmetric representation and the fixed location in the transcript of repetitive elements found on certain poly(A)$^+$ mRNA sets in Dictyostelium (Kindle and Firtel, 1979; Kimmel and Firtel, 1979) make them quite unlike the polyadenylated repeat transcripts in sea urchins.

Even if part of the length of these repeat-containing RNAs is capable of coding for protein, it seems certain that the regions in which repeat elements reside do not have a coding function. The sequence of one such repetitive region presented here demonstrates directly that it is non-translatable. The data of Chapter II indicate

that the sequences of most repeat elements probably lack open translational reading frames. Moreover, as discussed in the previous section, the high degree of polymorphism in genomic regions surrounding transcribed repeats is consistent with a non-coding character.

The very properties of polyadenylated repeat transcripts which seem to distinguish them from mRNAs are reminiscent of the characteristics of heterogeneous nuclear RNA (hnRNA). Thus, hnRNAs of many species and cell types (Lewin, 1980) consist of long transcripts (averaging 5-10 kb), in which repetitive sequences are known to be interspersed (e.g., Holmes and Bonner, 1974; Smith et al., 1974; Molloy et al., 1974; Fedoroff et al., 1977). A considerable fraction of these molecules is polyadenylated (Lewin, 1980). Moreover, the presence of intervening sequences in primary transcripts means that hnRNA contains substantial lengths of non-coding sequences not found in mature mRNA. Specific examples of repetitive sequences within intervening sequences have been identified (e.g., Ryffel et al., 1981). This attractive parallel breaks down in at least two respects, however, in the case of the repeat-containing transcripts of sea urchin eggs. These are known to be located primarily in the cytoplasm of the egg, rather than in the nucleus, and furthermore they are stable transcripts, synthesized in oogenesis and then stored, in contrast to hnRNAs, which turn over very rapidly. It should be noted that Anderson et al. (1982) have detected interspersed polyadenylated transcripts in manually enucleated Xenopus oocytes, thus explicitly demonstrating their cytoplasmic location.

How then can the existence of repeat-containing RNAs with the properties we have described be interpreted?

One proposal (Scheller et al., 1978; Davidson and Britten, 1979) has been that the repeat-containing transcripts in the nucleus fall into two classes: primary transcripts from structural genes, which would carry repetitive elements in their intervening or non-coding sequences, and regulatory RNAs derived from separate transcription

units, containing repeat sequences complementary to those in the structural gene transcripts. A model was described (Davidson and Britten, 1979) by which RNA-RNA duplexes formed between the repeat elements of these two types of transcript could play a role in tissue-specific regulation of RNA processing and thus mRNA expression. The relevance of interspersed RNAs in egg cytoplasm to this proposal was provided by the suggestion that the egg transcripts might be taken up into the nuclei of early blastomeres, there to participate in the regulatory interactions envisioned by the model.

A second proposal, recently articulated by Thomas et al. (1981), is that despite their stability and cytoplasmic location, the interspersed polyadenylated transcripts of sea urchin eggs are for the most part unprocessed or partially processed mRNA precursors. (This hypothesis requires that RNA processing and transport in oocyte nuclei are fundamentally different from these events in somatic nuclei.) Such unprocessed transcripts, it was suggested, might have regulatory roles in the embryo, or might instead be processed after fertilization to provide a flow of mature mRNAs. Much of the motivation for the latter idea came from the observation of Costantini et al. (1980) that a fraction of egg total RNA enriched in repeat-containing transcripts was also enriched for 70-80% of the single-copy sequence complexity of the egg. Earlier studies (Hough-Evans et al., 1977) had found the large majority of these single-copy sequences on early embryo polysomes. The implication (Costantini et al., 1980) is that a large proportion of the single copy sequences which constitute maternal mRNA in sea urchin embryos is covalently associated in the egg with repetitive sequences. However, this conclusion may have to be re-examined in light of the recent finding (Jacobs, Posakony, and Grula, unpublished) that SpG30, an embryo cDNA clone whose sequences were also shown to be enriched in the RNA fraction just described (Costantini et al., 1980), is of mitochondrial origin. Mitochondrial mRNAs would not be expected to carry repetitive nuclear sequences.

Observations described in this chapter are also relevant to the notion that the repeat-containing transcripts of the egg are mRNA precursors. The results shown in Figures 2 and 7 would seem to argue against this idea. Thus, while mRNA precursors might accumulate in the oocyte cytoplasm during oogenesis, it is difficult to understand why some of the same precursor species should achieve the high steady-state prevalence in later embryos implied by their strong hybridization signals on RNA gel blots [particularly of total poly(A)$^+$ RNA]. At the stages studied here (particularly gastrula), sea urchin embryo nuclei are actively synthesizing most or all of the mRNA found on the polysomes (Galau et al., 1977), and the hnRNA is turning over rapidly (Davidson, 1976). High mRNA precursor levels seem inconsistent with this state of affairs.

In the remainder of this section, two additional proposals for the origin, if not the function, of polyadenylated repeat-containing transcripts will be offered.

The first idea is that each transcript is initiated in a transcription unit that usually gives rise to a messenger RNA, and in general has already undergone whatever RNA splicing events are characteristic of the production of mRNA from that locus. However, the repeat-containing transcript would carry (compared to the mRNA) a long extension 3' to the protein-coding sequences, resulting from the failure of polyadenylation to occur at the site used in the mRNA. The fact that many repeat-containing transcripts are discrete and polyadenylated would be due to poly(A) addition at a downstream polyadenylation site, perhaps a fortuitous one. This proposal requires that the RNA polymerase continue past the "normal" polyadenylation site in producing such a transcript. The observation by Fraser et al. (1979) of nuclear transcripts extending beyond the last poly(A) addition site of the major adenovirus late transcription unit provides an appropriate precedent. [Indeed, the work of these and other investigators (Nevins and Darnell, 1978; Nunberg et al., 1980; Hofer and Darnell, 1981) suggests that cleavage of the primary transcript, followed by polyadenylation, is the mechanism

by which eukaryotic mRNA ends are defined, rather than polymerase termination.] The interspersed nature of a transcript produced as described would reflect the interspersion of repetitive and single-copy sequences in the DNA downstream from the mRNA coding region. (The bypassing of the usual polyadenylation site could be either random or subject to regulation.) Having undergone normal splicing, such a transcript could conceivably be exported normally from the nucleus and perhaps even translated. A testable prediction of this model is that protein-coding sequences, but not repeat sequences, will be found close to the 5' ends of interspersed RNAs.

An important impetus for the foregoing came from observations in the laboratories of MacGregor and Gall of transcription of satellite DNA sequences in the lampbrush chromosomes of newt oocytes (Varley et al., 1980; Diaz et al., 1981). It was suggested by Varley et al. (1980) that such transcription might be the result of failure of transcription termination in a gene or genes, causing polymerase to "read through" into adjacent satellite sequences. Diaz et al. (1981) were able to provide very strong evidence for just such read-through transcripts, initiating at histone gene promoters and continuing through flanking satellite DNA. MacGregor's group (MacGregor et al., 1981) has also presented evidence of read-through transcription of ribosomal RNA genes by polymerase II or III.

A second, much more speculative, proposal which will be briefly considered is that repeat-containing transcripts originate at fortuitous promoters lying within transcriptionally active domains of the chromosome. Again, many of these might become polyadenylated by virtue of their containing fortuitous or bona fide polyadenylation signals. The fundamental idea here is that a chromosomal region which has become permissive for transcription in order to activate an essential gene may contain fortuitous or vestigial promoters, initiation at which could give rise to interspersed transcripts as a reflection of the local DNA sequence organization. One consequence of this model is a possible explanation for developmental modulation

of repeat sequence transcription, since the appearance of new interspersed transcripts would follow the temporal pattern of chromosomal domain activation used by the embryo to produce new mRNAs.

The various ideas considered in this section are by no means mutually exclusive, and each of them may be correct for some fraction of repeat-containing transcripts. Clearly, the detailed analysis of individual transcripts and transcription units—in their entirety—is essential. The cDNA clones described in this chapter provide a starting point.

Whatever the significance of interspersed polyadenylated transcripts in sea urchin eggs, evidence presented in this chapter makes it likely that the exact set of RNA molecules which carry a given repeat sequence is of relatively little importance. Thus, the sizes and relative prevalences of these transcripts have not been conserved in the 10-20 million years (Durham, 1966) since the divergence of lineages leading to S. purpuratus and S. franciscanus. Nevertheless, the overall concentration in egg RNA of transcripts complementary to given repeats is much more stable phylogenetically (this thesis, Chapter III), and this may be a functionally important parameter.

REFERENCES

Alexandraki, D., and Ruderman, J. V. (1981). Mol. Cell. Biol. **1**:1125-1137.

Anderson, D. M., Richter, J. D., Chamberlin, M. E., Price, D. H., Britten, R. J., Smith, L. D., and Davidson, E. H. (1982). J. Mol. Biol. **155**, in press.

Aviv, H., and Leder, P. (1972). Proc. Natl. Acad. Sci. USA **69**:1408-1412.

Bailey, J. M., and Davidson, N. (1976). Analyt. Biochem. **70**:75-85.

Britten, R. J., Graham, D. E., and Neufeld, B. R. (1974). In Methods in Enzymology (Grossman, L. & Moldave, K., eds.), vol. 29E, pp. 363-406, Academic Press, New York.

Britten, R. J., Cetta, A., and Davidson, E. H. (1978). Cell **15**:1175-1186.

Brosius, J., Palmer, M. L., Kennedy, P. J., and Noller, H. F. (1978). Proc. Natl. Acad. Sci. USA **75**:4801-4805.

Brosius, J., Dull, T. J., and Noller, H. F. (1980). Proc. Natl. Acad. Sci. USA **77**:201-204

Calabretta, B., Robberson, D. L., Maizel, A. L., and Saunders, G. F. (1981). Proc. Natl. Acad. Sci. USA **78**:6003-6007.

Casey, J., and Davidson, N. (1977). Nucleic Acids Res. **4**:1539-1552.

Childs, G., Maxson, R., and Kedes, L. H. (1979). Dev. Biol. **73**:153-173.

Cohen, S. N., Chang, A. C. Y., and Hsu, L. (1972). Proc. Natl. Acad. Sci. USA **69**:2110-2114.

Costantini, F. D., Scheller, R. H., Britten, R. J., and Davidson, E. H. (1978). Cell **15**:173-187.

Costantini, F. D., Britten, R. J., and Davidson, E. H. (1980). Nature **287**:111-117.

Crain, W. R., Jr., Durica, D. S., and Van Doren, K. (1981). Mol. Cell. Biol. **1**:711-720.

Crampton, J. M., Davies, K. E., and Knapp, T. F. (1981). Nucleic Acids Res **9**:3821-3834.

Davidson, E. H. (1976). Gene Activity in Early Development, 2nd ed., Academic Press, New York.

Davidson, E. H., and Britten, R. J. (1979). Science **204**:1052-1059.

Davidson, E. H. (1976). Gene Activity in Early Development, 2nd ed., Academic Press, New York.

Davidson, E. H., and Britten, R. J. (1979). Science **204**:1052-1059.

Davidson, E. H., and Posakony, J. W. (1982). Nature, in press.

Dawid, I. B., Long, E. O., DiNocera, P. P., and Pardue, M. L. (1981). Cell **25**:399-408.

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

Diaz, M. O., Barsacchi-Pilone, G., Mahon, K. A., and Gall, J. G. (1981). Cell **24**:649-659.

Durham, J. W. (1966). In Treatise on Invertebrate Paleontology (U) Echinodermata (Moore, R. C., ed.), vol. 3, part 1, pp. 270-295, The Geological Society of America and The University of Kansas Press, New York.

Durica, D. S., Schloss, J. A., and Crain, W. R., Jr. (1980). Proc. Natl. Acad. Sci. USA **77**:5683-5687.

Ernst, S. G., Hough-Evans, B. R., Britten, R. J., and Davidson, E. H. (1980). Dev. Biol. **79**:119-127.

Fedoroff, N., Wellauer, P. K., and Wall, R. (1977). Cell **10**:597-610.

Flytzanis, C. N., Brandhorst, B. P., Britten, R. J., and Davidson, E. H. (1982). Dev. Biol. **90**, in press.

Fraser, N. W., Nevins, J. R., Ziff, E., and Darnell, J. E., Jr. (1979). J. Mol. Biol. **129**:643-656.

Galau, G. A., Lipson, E. D., Britten, R. J., and Davidson, E. H. (1977). Cell **10**:415-432.

Glass, J., and Wertz, G. W. (1980). Nucleic Acids Res. **8**:5739-5751.

Glisin, V., Crkvenjakov, R., and Byus, C. (1974). Biochemistry **13**:2633-2637.

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

Hanahan, D., and Meselson, M. (1980). Gene **10**:63-67.

Hofer, E., and Darnell, J. E., Jr. (1981). Cell **23**:585-593.

Holmes, D. S., and Bonner, J. (1974). Proc. Natl. Acad. Sci. USA **71**:1108-1112.

Hough-Evans, B. R., Wold, B. J., Ernst, S. G., Britten, R. J., and Davidson, E. H. (1977). Dev. Biol. **60**:258-277.

Hynes, R. O., and Gross, P. R. (1970). Dev. Biol. **21**:383-402.

Kimmel, A. R., and Firtel, R. A. (1979). Cell **16**:787-796.

Kindle, K. L., and Firtel, R. A. (1979). Nucleic Acids Res. **6**:2403-2422.

Kitamura, N., Semler, B. L., Rothberg, P. G., Larsen, G. R., Adler, C. J., Dorner, A. J., Emini, E. A., Hanecak, R., Lee, J. J., van der Werf, S., Anderson, C. W., and Wimmer, E. (1981). Nature **291**:547-553.

Klein, W. H., Thomas, T. L., Lai, C., Scheller, R. H., Britten, R. J., and Davidson, E. H. (1978). Cell **14**:889-900.

Lehrach, H., Diamond, D., Wozney, J. M., and Boedtker, H. (1977). Biochemistry **16**:4743-4751.

Lewin, B. (1980). Gene Expression, vol. 2, 2nd ed., John Wiley & Sons, New York.

MacGregor, H. C., Varley, J. M., and Morgan, G. T. (1981). In International Cell Biology 1980-1981 (Schweiger, H. G., ed.), pp. 33-46, Springer-Verlag, Berlin.

Maxam, A. M., and Gilbert, W. (1977). Proc. Natl. Acad. Sci. USA **74**:560-564.

Maxam, A. M., and Gilbert, W. (1980). In Methods in Enzymology (Grossman, L., and Moldave, K., eds.), vol. 65, pp. 499-560, Academic Press, New York.

Molloy, G. R., Jelinek, W., Salditt, M., and Darnell, J. E. (1974). Cell **1**:43-53.

Moore, G. P., Scheller, R. H., Davidson, E. H., and Britten, R. J. (1978). Cell **15**:649-660.

Moore, G. P., Pearson, W. R., Davidson, E. H., and Britten, R. J. (1981). Chromosoma (Berl.) **84**:19-32.

Nevins, J. R., and Darnell, J. E., Jr. (1978). Cell **15**:1477-1493.

Norgard, M. V., Emigholz, K., and Monahan, J. J. (1979). J. Bacteriol. **138**:270-272.

Nunberg, J. H., Kaufman, R. J., Chang, A. C. Y., Cohen, S. N., and Schimke, R. T. (1980). Cell **19**:355-364.

Rave, N., Crkvenjakov, R., and Boedtker, H. (1979). Nucleic Acids Res. **6**:3559-3567.

Rowekamp, W., and Firtel, R. A. (1980). Dev. Biol. **79**:409-418.

Ryffel, G. U., Muellener, D. B., Wyler, T., Wahli, W., and Weber, R. (1981). Nature **291**:429-431.

Scheller, R. H., Thomas, T. L., Lee, A. S., Klein, W. H., Niles, W. D., Britten, R. J., and Davidson, E. H. (1977). Science **196**:197-200.

Scheller, R. H., Costantini, F. D., Kozlowski, M. R., Britten, R. J., and Davidson, E. H. (1978). Cell **15**:189-203.

Scheller, R. H., Anderson, D. A., Posakony, J. W., McAllister, L. B., Britten, R. J., and Davidson, E. H. (1981a). J. Mol. Biol. **149**:15-39.

Scheller, R. H., McAllister, L. B., Crain, W. R., Jr., Durica, D. S., Posakony, J. W., Thomas, T. L., Britten, R. J., and Davidson, E. H. (1981b). Mol. Cell. Biol. **1**:609-628.

Smith, M. J., Hough, B. R., Chamberlin, M. E., and Davidson, E. H. (1974). J. Mol. Biol. **85**:103-126.

Southern, E. M. (1975). J. Mol. Biol. **98**:503-517.

Sutcliffe, J. G. (1979). Cold Spring Harbor Symp. Quant. Biol. **43**:77-90.

Taylor, J. M., Illmensee, R., and Summers, J. (1976). Biochim. Biophys. Acta **442**:324-330.

Thomas, T. L., Posakony, J. W., Anderson, D. M., Britten, R. J., and Davidson, E. H. (1981). Chromosoma (Berl.) **84**:319-335.

Varley, J. M., MacGregor, H. C., and Erba, H. P. (1980). Nature **283**:686-688.

Wilt, F. H. (1977). Cell **11**:673-681.

Zuker, C., and Lodish, H. F. (1981). Proc. Natl. Acad. Sci. USA **78**:5386-5390.

Table I

Summary of genome blot and clone blot experiments with selected fragments of the cDNA clones

Source Clone[a]	Fragment[a]	Genome Blot[b]	Clone Blot[c]
p2109A-2	2A	SC	NS
	2AH	R	S
	2HH	SC	NS
p2109A-10	10A	R	S
	10HB	SC	NS
p2109A-16	16A	R	S
	16AP	SC	NS
	16PB	SC	NS
p2109A-18	18H1	R	S
	18H2	R	S
	18AR2	SC	NS
	18AR1	SC	NS
p2137-1	1HT	SC	NS
	1PP	R	S
p2137-5	5A1	R	NS
	5A2	R	NS
	5AP	R	S
p2137-7	7HR	R	S
	7PD	R	NS

Table I (continued)

p2137-9	9BP2	R	NS
	9BP1	R	S

[a]See Figure 3.

[b]See Figure 5A and B. SC = single-copy; R = repetitive.

[c]See Figure 5D and E. S = fragment sequences are shared with the other three clones in the set. NS = fragment sequences not shared; unique to source clone.

Table II

Summary of data[a] on polymorphism of genomic regions surrounding transcribed repeats

Probe Fragment[b]	Digest[c]	Minimum Number of Alleles[d]	Fraction Heterozygous[e]
2A	Eco RI	6	4/6
	Hind III	6	5/6
10HB	Eco RI	9	6/6
	Hind III	8	6/6
16PB	Eco RI	7	5/6
	Hind III	9	6/6
18AR1	Eco RI	9	5/6
	Hind III	7	4/6
1HT/1D	Eco RI	10	6/6
	Hind III	7	5/6

[a]For representative examples, see Figure 6A-D.

[b]See Figure 3.

[c]Of genomic DNAs.

[d]Number of bands of clearly distinguishable size in a blot of six individual DNAs (12 haploid genomes).

[e]Number of heterozygous (two clearly distinguishable bands in genome blot) individuals out of six tested.

Table III

Sizes of polyadenylated transcripts complementary to selected fragments of the cDNA clones

Probe Fragment[a]	Transcript Size(s) (kb)[b]
10HB	4.1
	7.6 (gastrula only)
16PB	9.2
18AR1	6.2
	3.2 (blastula only)
5A2	4.5 (egg and blastula)

[a]See Figure 3.

[b]Determined from RNA markers of known size (see Materials and Methods). If not otherwise indicated, transcript appears at all developmental stages studied. The size of the transcript complementary to fragment 16PB was determined by extrapolation.

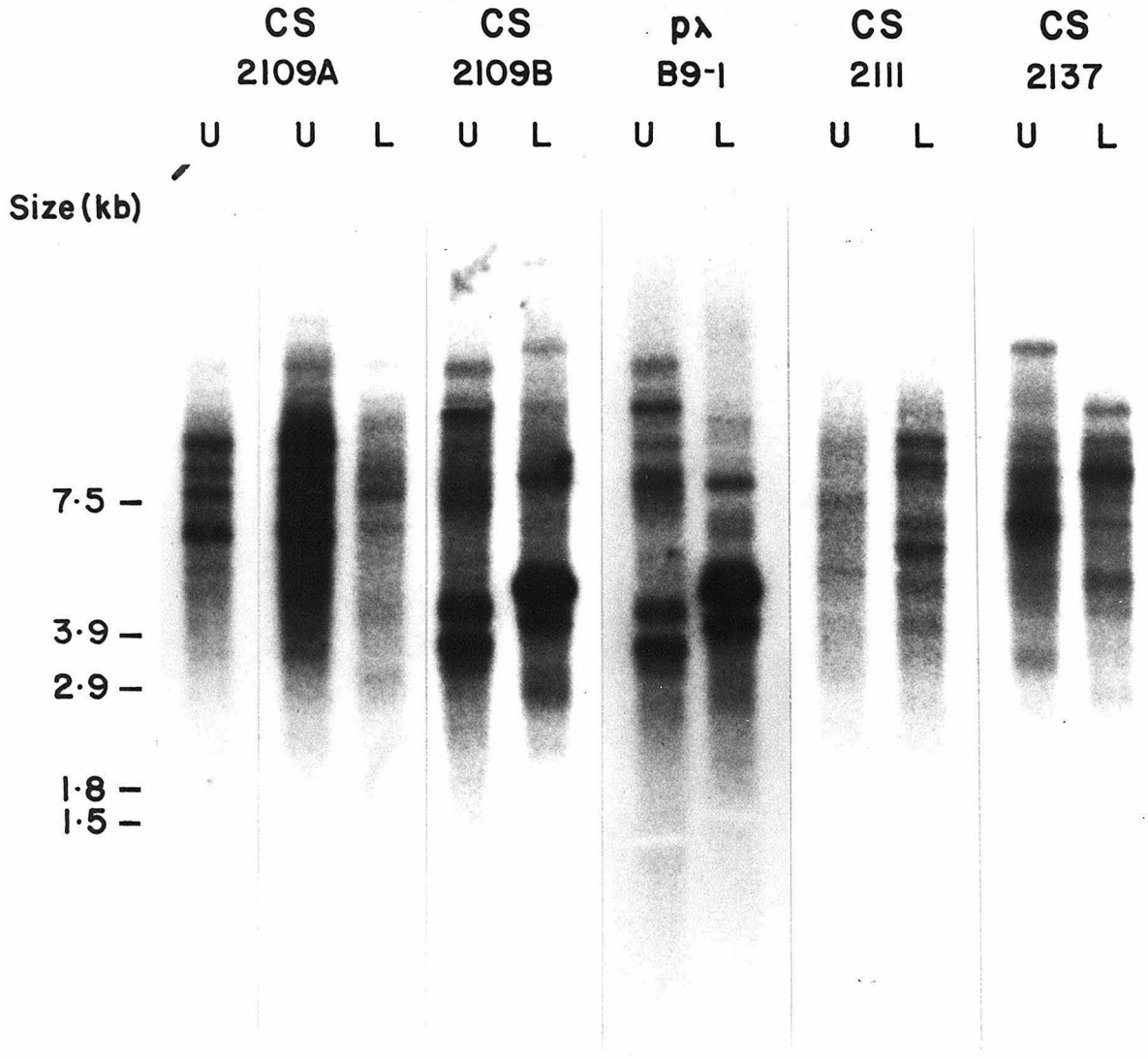
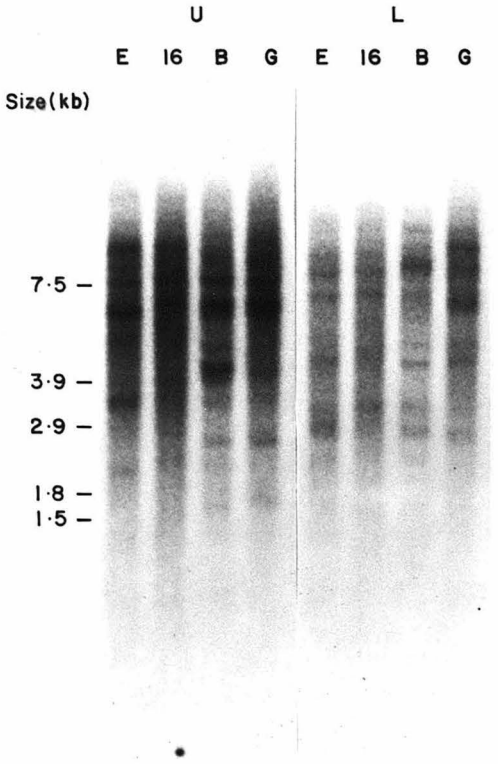

CS
2109A
CS
2109B
pλ
B9-1
CS
2111
CS
2137
U U L U L U L U L U L
Size (kb)
7·5 –
3·9 –
2·9 –
1·8 –
1·5 –

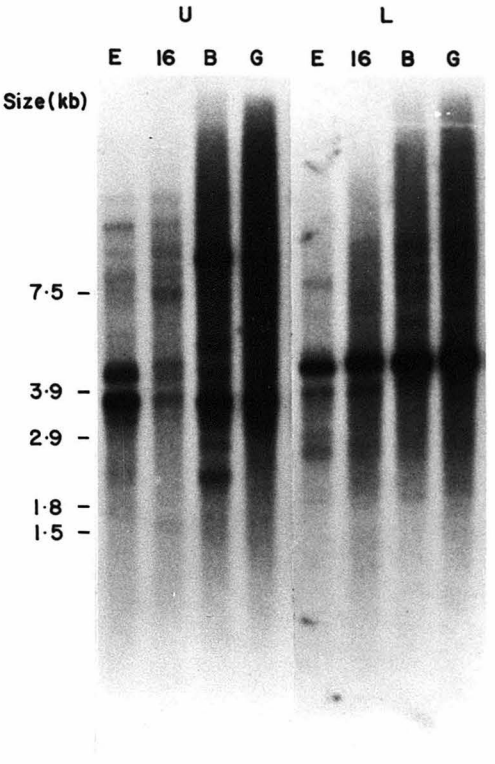
Figure 1

Gel blots of S. purpuratus egg poly(A)$^+$ RNA hybridized with the separated strands of several cloned repetitive sequence elements (Klein et al., 1978; this thesis, Chapter II). RNA (1 μg/lane) was denatured, electrophoresed, and transferred to a nitrocellulose filter as described in Materials and Methods. Hybridization was carried out under 20% formamide criterion conditions (see Materials and Methods). The strand designations U (upper) and L (lower) refer to the relative positions of the strands on neutral polyacrylamide strand separation gels (see Materials and Methods). The plasmid clone pλB9-1 contains a member of the 2109B repeat family. It is a subclone of a 600-nt Hae III fragment isolated from λ2109B-9 (this thesis, Chapter I), inserted into pBR322 by means of Eco RI linkers. As shown in Figure 2(b) of Scheller et al. (1981a), this particular member of the 2109B family is substantially divergent from the cloned element CS2109B. The size scale shows the migration of RNA markers (see Materials and Methods) of known size, electrophoresed in parallel lanes. The leftmost lane shows a lower exposure of the filter hybridized to the CS2109A U strand. Otherwise, exposures shown for the two strands of each pair were the same.

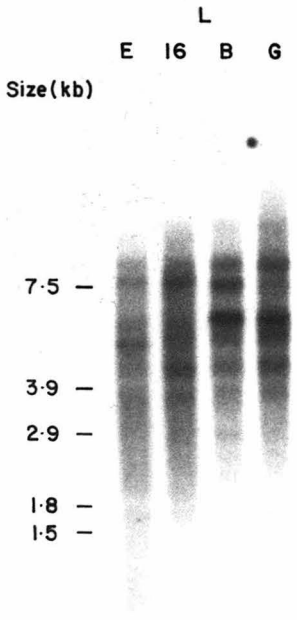
CS2109A



CS2109B



CS2111



CS2137

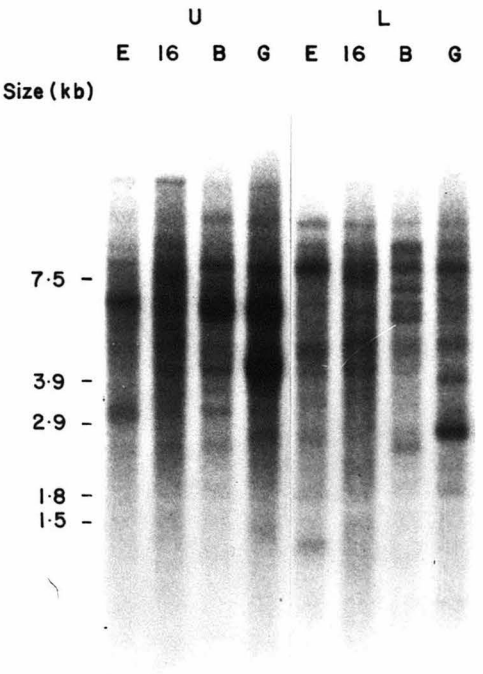
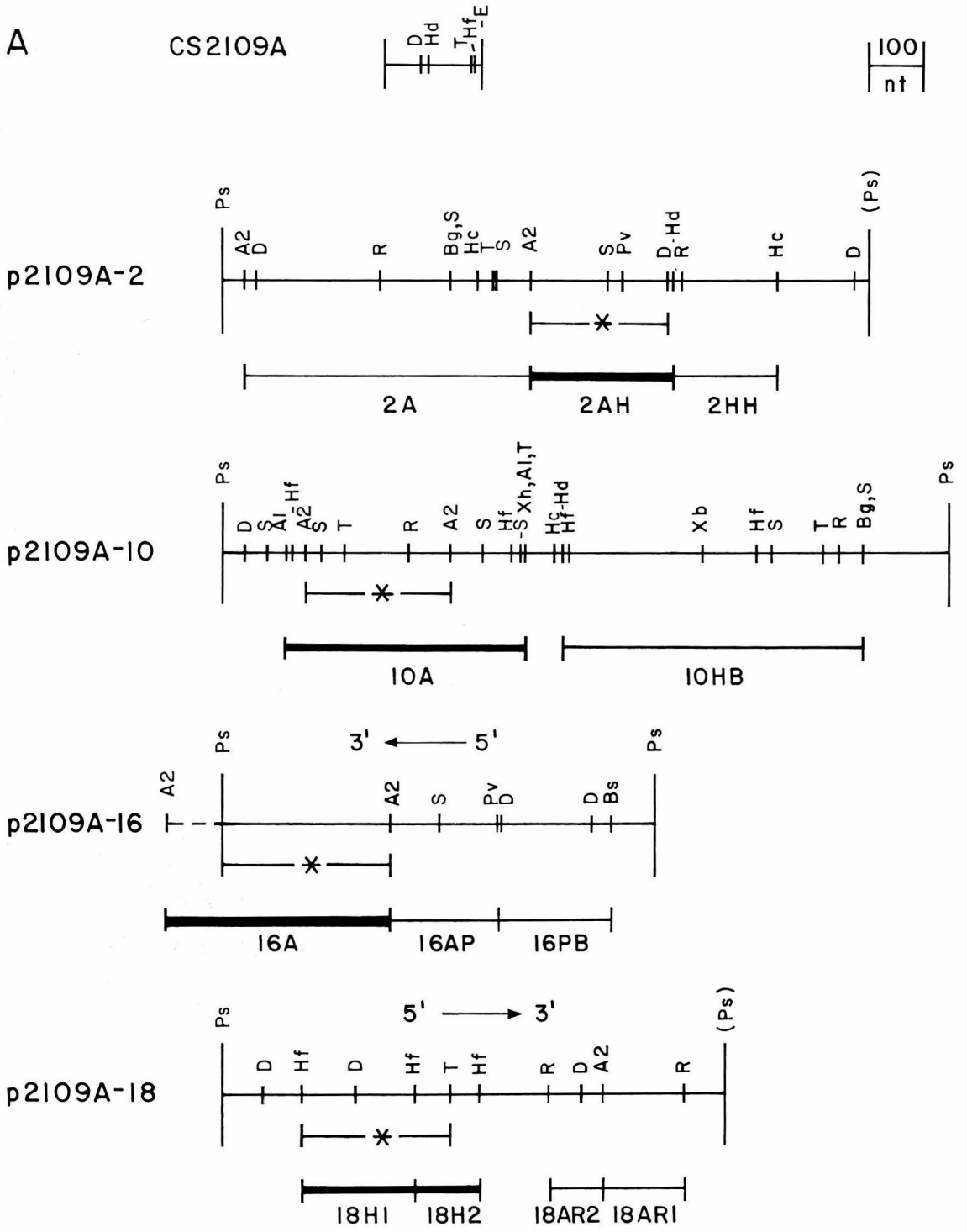


Figure 2

Gel blots of total poly(A)$^+$ RNA's isolated from different developmental stages of S. purpuratus, hybridized with the separated strands of several cloned repeat elements. E, egg; 16, 16-cell stage (5 h post-fertilization at 15°C); B, hatched (swimming) blastula stage (18 h p.f.); G, mid-gastrula stage (40 h p.f.). RNA (2 µg/lane) was denatured, electrophoresed, and transferred to nitrocellulose filters as described in Materials and Methods. Hybridization was carried out under 20% formamide criterion conditions (see Materials and Methods). Strand designations are as in Figure 1. Size scales show the migration of RNA markers of known size (see Materials and Methods), electrophoresed in parallel lanes.


A
CS2109A
D
Hd
T
Hf
E
100
nt
p2109A-2
Ps
A2
D
R
Bg,S
Hc
T
S
A2
S
Pv
D
R
Hd
Hc
D
(Ps)
2A
2AH
2HH
p2109A-10
Ps
D
S
Al
A2
Hf
S
T
R
A2
S
Hf
S
Xh,Al,T
Hc
Hf
Hd
Xb
Hf
S
T
R
Bg,S
Ps
10A
10HB
p2109A-16
Ps
3'
5'
Ps
A2
A2
S
Pv
D
D
Bs
16A
16AP
16PB
p2109A-18
Ps
5'
3'
(Ps)
D
Hf
D
Hf
T
Hf
R
D
A2
R
18H1
18H2
18AR2
18AR1

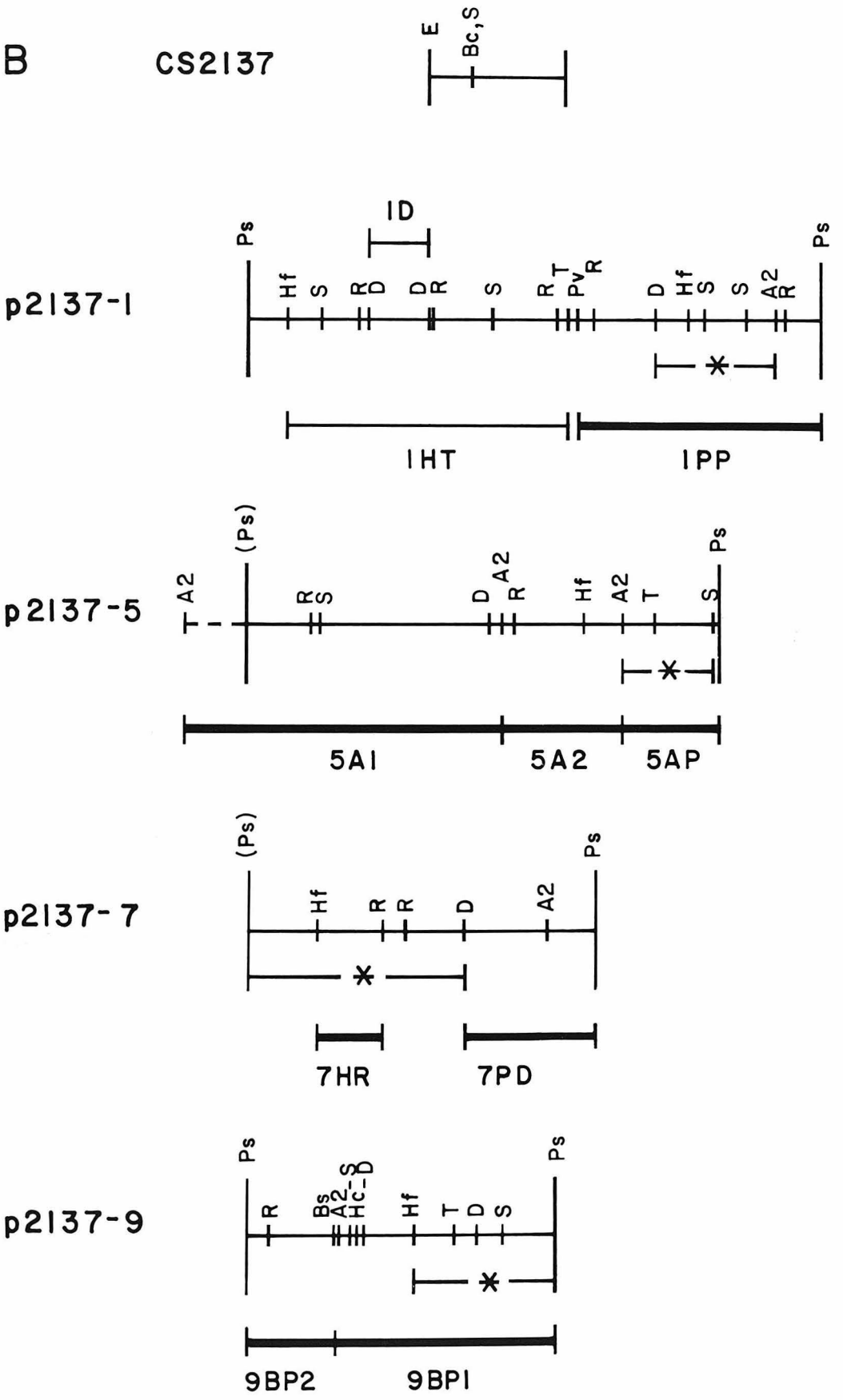

B
CS2137
E
Bc,S
p2137-1
1D
Ps
Hf
S
R
D
D
R
S
R
T
Pv
R
D
Hf
S
S
A2
R
Ps
1HT
1PP
p2137-5
T
A2
(Ps)
R
S
D
A2
R
Hf
A2
T
S
Ps
5A1
5A2
5AP
p2137-7
(Ps)
Hf
R
R
D
A2
Ps
7HR
7PD
p2137-9
Ps
R
Bs
A2-S
Hc-D
Hf
T
D
S
Ps
9BP2
9BP1

Figure 3

Restriction maps of cDNA clones of egg poly(A)$^+$ RNA, isolated by homology to the cloned repetitive elements CS2109A (A) and CS2137 (B). Restriction maps of these elements are also shown for comparison. Details of cDNA clone preparation and screening are given in Materials and Methods. Restriction site designations are as follows: A1, Ava I; A2, Ava II; Bc, Bcl I; Bg, Bgl II; Bs, BstE II; D, Dde I; E, Eco RI; Hc, Hinc II; Hd, Hind III; Hf, Hinf I; Ps, Pst I; Pv, Pvu II; R, Rsa I; S, Sau 3AI; T, Taq I; Xb, Xba I; Xh, Xho I. The designation "(Ps)" refers to vector Pst I sites not regenerated during cloning (see Materials and Methods); thus, the enzyme does not cut at these sites. All clones were treated with a set of 24 different restriction enzymes (see Materials and Methods); if a particular site designation does not appear on the map of a given clone, this signifies that the cDNA insert lacks a site for the corresponding enzyme. The minimal region of each cDNA clone to which homology to the CS repeat probe could be localized by blot hybridization is indicated by an asterisk below the restriction map of the clone. Restriction fragments used as probes in gel blot hybridizations are indicated below the map of the source clone. Fragment designations refer to the number of the source clone and the restriction enzyme(s) used to generate the fragment. Restriction fragments represented by thin lines were found to be single-copy in genomic DNA (see text). Fragments represented by thick lines were found to contain repetitive sequences. Arrows above the restriction maps of p2109A-16 and p2109A-18 (A) indicate their 5'-to-3' orientation in RNA, determined by experiments shown in Figure 8.

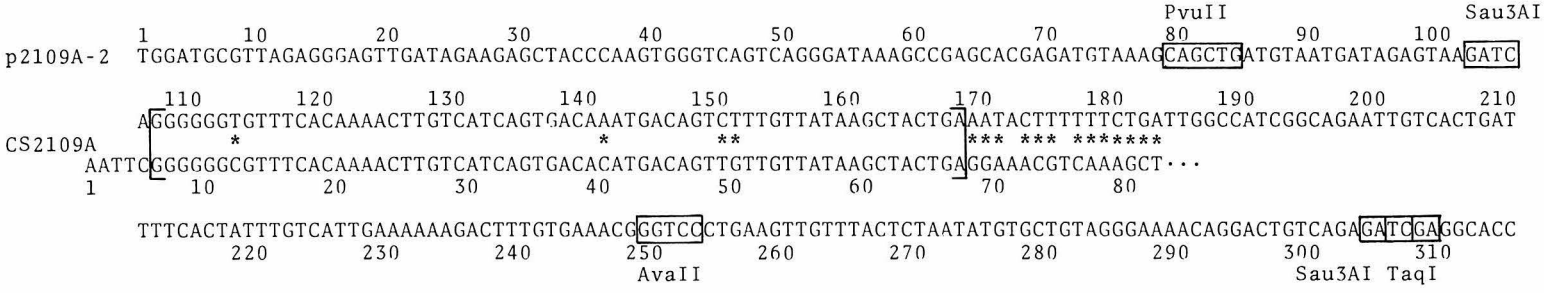
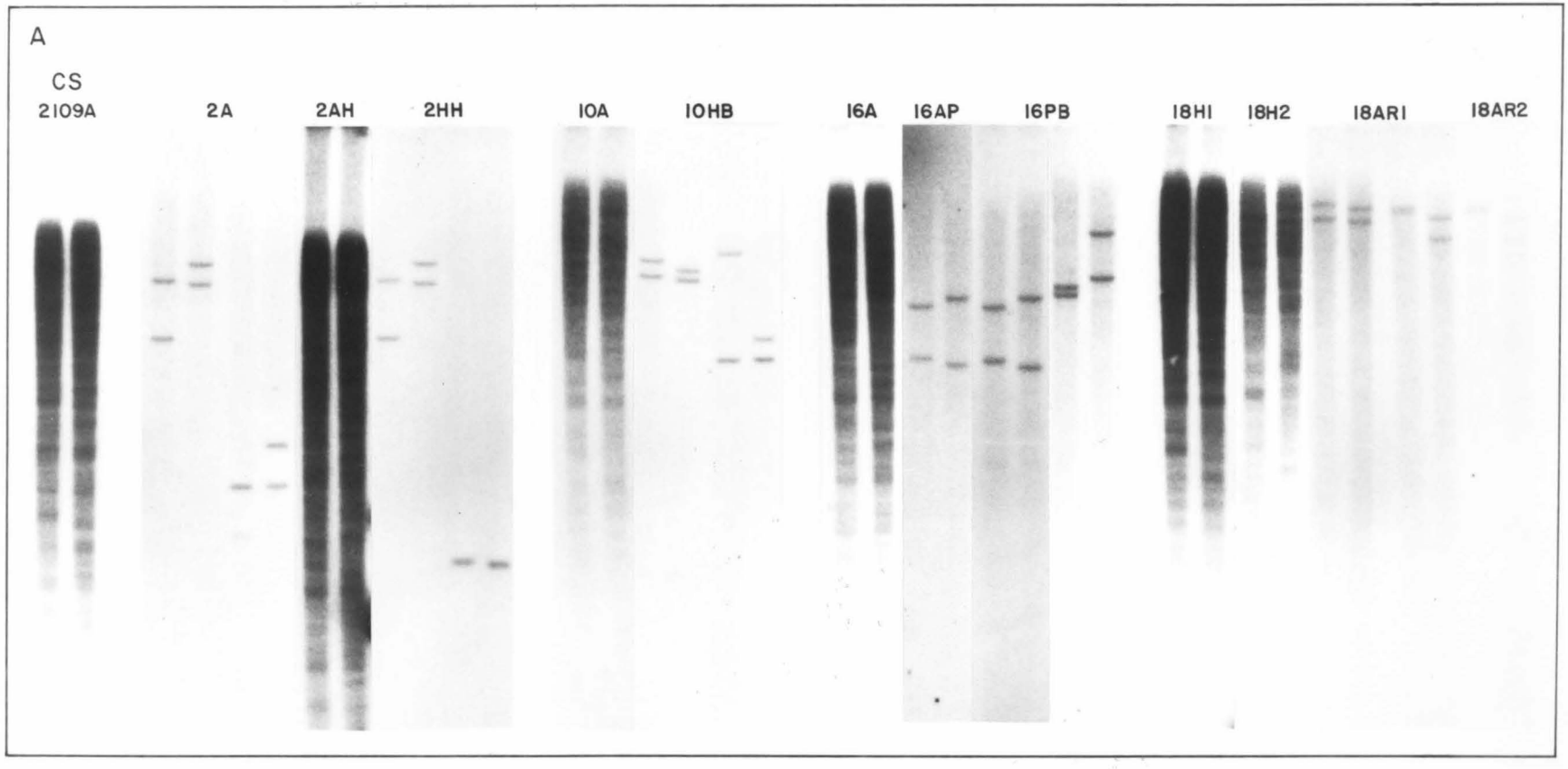
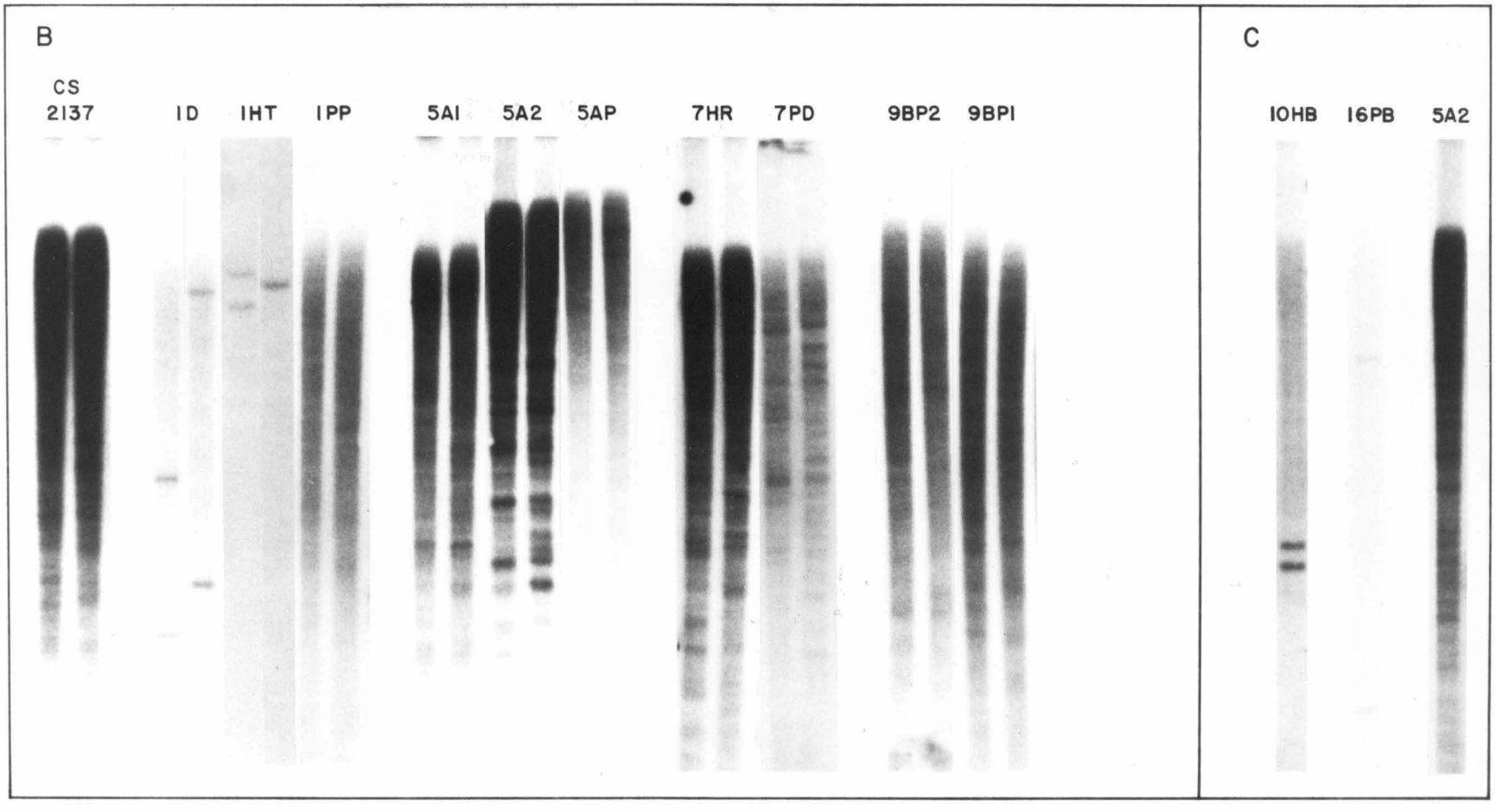

PvuII
Sau3AI
p2109A-2
CS2109A
AvaII
Sau3AI TaqI

Figure 4

Nucleotide sequence of the region in p2109A-2 containing homology to the cloned repeat element CS2109A. For comparison, the sequence of part of CS2109A (this thesis, Chapter II) is shown aligned below the appropriate region of the p2109A-2 sequence. Asterisks indicate base differences, and the region of homology between the two sequences is enclosed in brackets. The sequences are numbered from their 5' ends. Previously mapped restriction sites in p2109A-2 confirmed by the sequencing are enclosed in boxes. (The left-to-right orientation of this figure is opposite to that of the restriction map of p2109A-2 shown in Figure 3A.) The sequence shown was obtained by the method of Maxam and Gilbert (1980) from a 360-nt Hinc II-Hind III fragment (see Figure 3A) labeled asymmetrically at the Hind III end by Klenow polymerase (see Materials and Methods). The sequence from nucleotides 88 to 234 was confirmed by sequencing a 170-nt Ava II-Pvu II fragment (see Figure 3A) labeled asymmetrically at the Ava II end.


A
CS
2109A
2A
2AH
2HH
10A
10HB
16A
16AP
16PB
18H1
18H2
18AR1
18AR2


B
CS
2137
1D
1HT
1PP
5A1
5A2
5AP
7HR
7PD
9BP2
9BP1
C
10HB
16PB
5A2

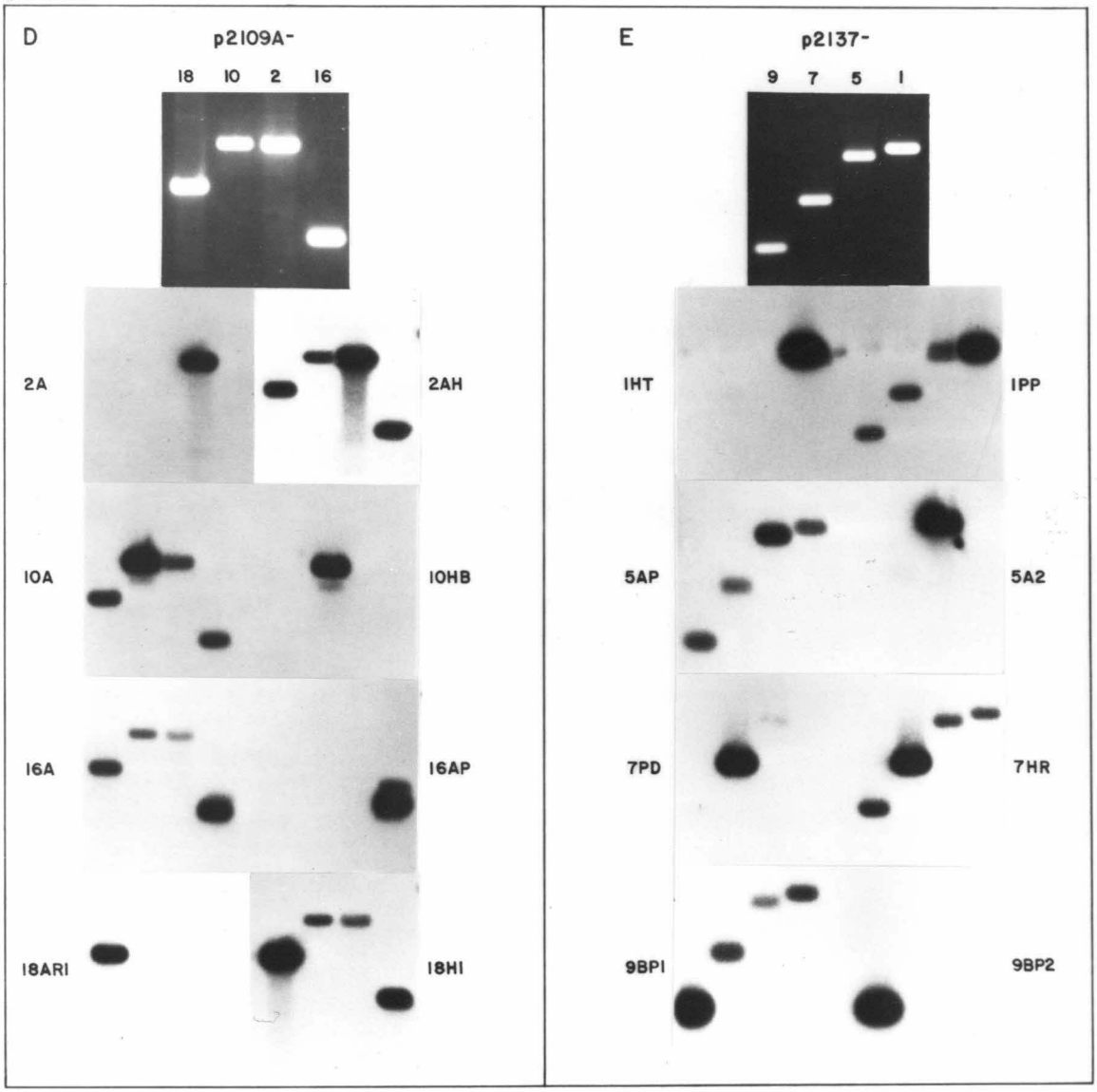

D
p2109A-
18
10
2
16
2A
2AH
10A
10HB
16A
16AP
18ARI
18HI
E
p2137-
9
7
5
1
1HT
1PP
5AP
5A2
7PD
7HR
9BP1
9BP2

Figure 5

DNA gel blot hybridizations using selected restriction fragments from the cDNA clones (see Figure 3).

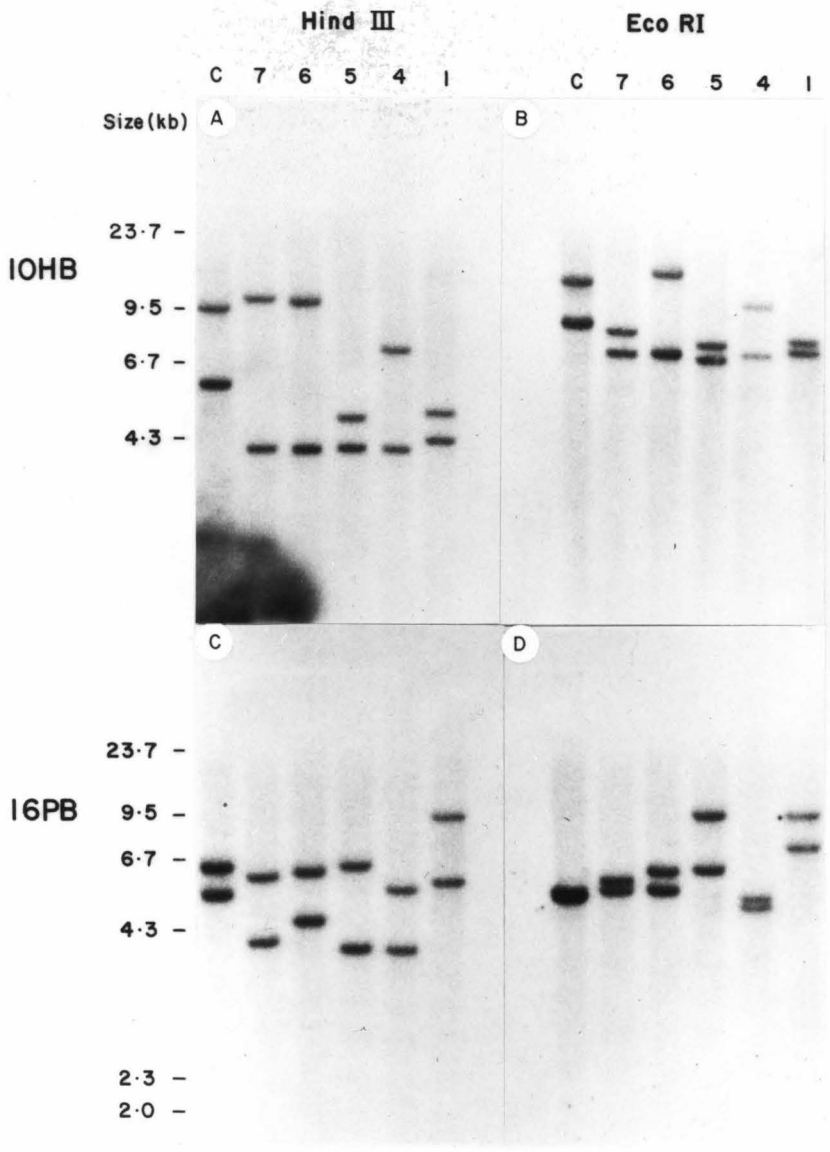
A and B: Restriction enzyme digests of whole genomic DNA of two S. purpuratus individuals reacted on gel blots with labeled restriction fragments from p2109A cDNA clones (A) or p2137 cDNA clones (B). The shorthand designation (see Figure 3) of the fragment used as a probe appears above the corresponding autoradiogram. In each pair of lanes the DNA of individual No. 7 is on the left, that of individual No. 5 on the right, digested with the same enzyme. From left to right, the restriction enzyme digests of genomic DNA are as follows: (A) CS2109A, Eco RI; 2A, Eco RI, Hind III; 2AH, Hind III; 2HH, Eco RI, Hind III; 10A, Eco RI; 10HB, Eco RI, Hind III; 16A, Hind III; 16AP, Hind III; 16PB, Hind III, Eco RI; 18H1, Eco RI; 18H2, Eco RI; 18AR1, Eco RI, Hind III; 18AR2, Hind III. (B) CS2137, Eco RI; 1D, Eco RI; 1HT, Hind III; 1PP, Hind III; 5A1, Eco RI; 5A2, Eco RI; 5AP, Eco RI; 7HR, Eco RI; 7PD, Eco RI; 9BP2, Eco RI; 9BP1, Hind III.

C: Restriction enzyme digests of whole genomic DNA of a single S. franciscanus individual reacted on gel blots with labeled restriction fragments from the cDNA clones. The shorthand designation (see Figure 3) of the fragment used as a probe appears above the corresponding autoradiogram. Restriction enzyme digests of the genomic DNA are as follows: 10HB, Hind III; 16PB, Hind III; 5A2, Eco RI.

D and E: Insert fragments of the p2109A cDNA clones (D) or the p2137 cDNA clones (E) reacted on gel blots with labeled restriction fragments of the cDNA clones. Only the region of each gel or autoradiogram containing the insert fragments is shown. The insert fragments were released from vector DNA by digestion with Pst I. In cDNA clones lacking a Pst I site at one end of the insert (Figure 3), a second restriction enzyme, cutting at a site in the adjacent vector DNA, was used to release the insert fragment. Thus, insert fragments of p2109A-2 and p2109A-18 were released

Figure 5 (continued)

by a Pst I + Pvu I double digestion; inserts of p2137-5 and p2137-7 by Pst I + Bgl I. At the top of each set of autoradiograms is shown a photograph of a representative gel, stained with ethidium bromide and illuminated with ultraviolet light. The insert fragment in each lane is identified with the number of the corresponding cDNA clone. Below are autoradiograms of gel blots reacted with the labeled cDNA clone restriction fragments (Figure 3) indicated on each side.


Hind III
Eco RI
C 7 6 5 4 I
C 7 6 5 4 I
Size(kb)
A
B
C
D
10HB
16PB
23·7
9·5
6·7
4·3
2·3
2·0

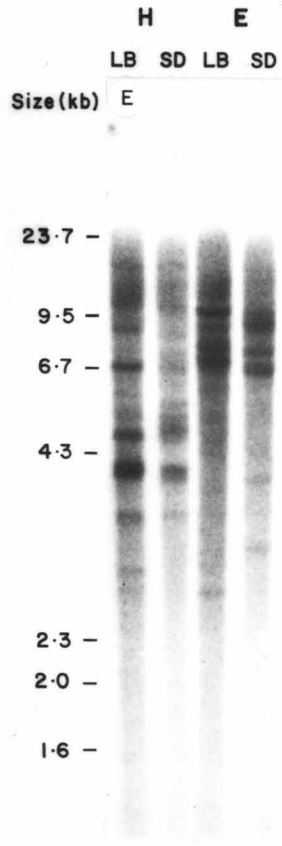

H
E
LB SD LB SD
Size(kb)
E
23·7
9·5
6·7
4·3
2·3
2·0
1·6

Figure 6

A - D: Restriction enzyme digests of whole genomic DNA of six S. purpuratus individuals, reacted on gel blots with single-copy fragments from two of the cDNA clones (Figure 3): 10HB from p2109A-10 (A and B) and 16PB from p2109A-16 (C and D). Restriction enzyme digests of genomic DNA are as indicated. Each lane is labeled with the designation (1, 4, 5, 6, 7, or C) of the appropriate individual.

E: Restriction enzyme digests of mixtures of equal amounts of whole genomic DNA of many S. purpuratus individuals, reacted on gel blots with the single-copy fragment 10HB from p2109A-10 (Figure 3). "LB" designates a mixture of DNA's from 30 individuals collected near Laguna Beach, California; "SD" designates a mixture of DNA's from 20 individuals collected near San Diego (Point Loma), California. Restriction enzyme digests of the DNA mixtures were Hind III (H) and Eco RI (E).

A - E: Size scales show the migration of fragments of Hind III - digested λ DNA.

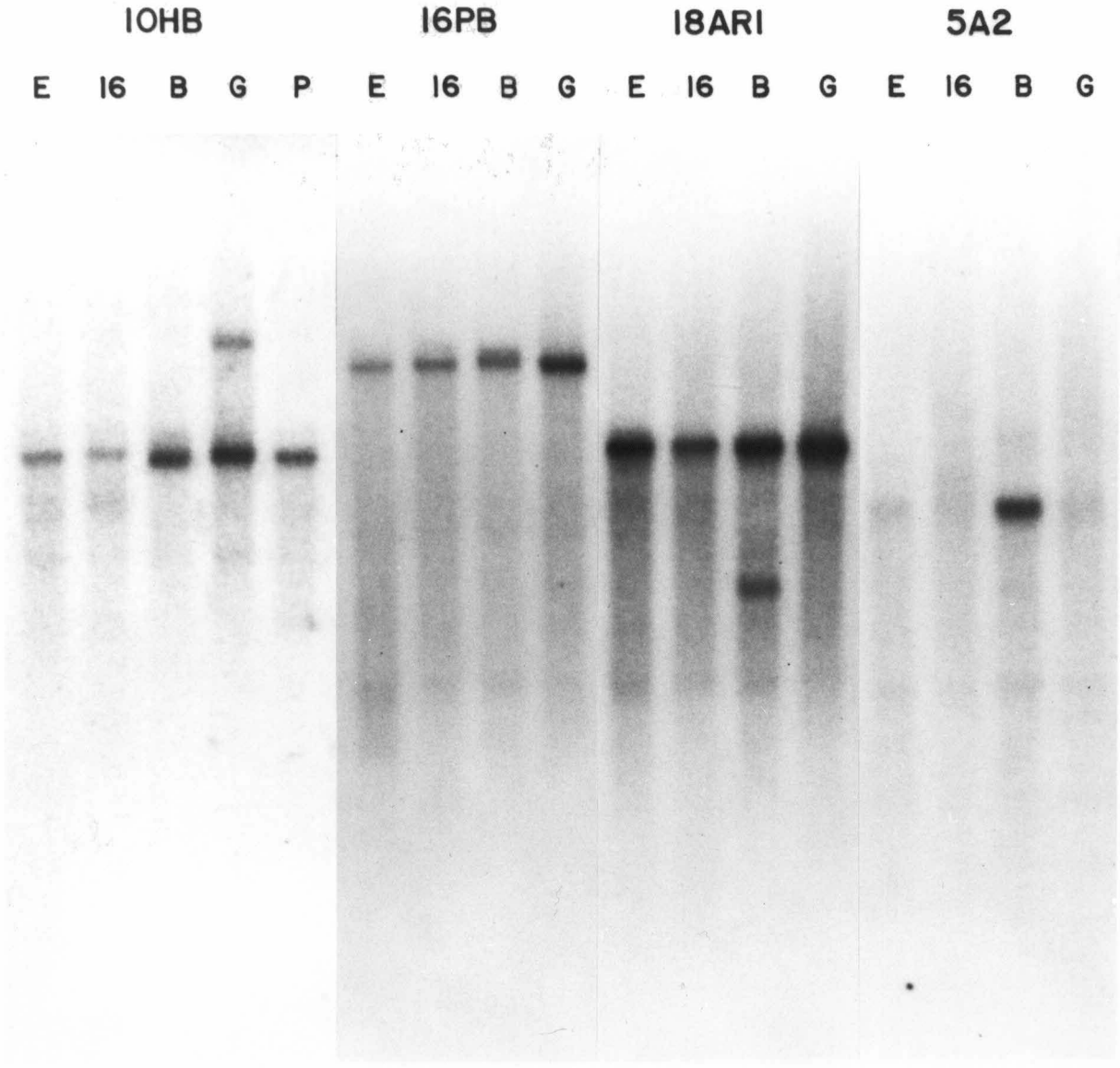
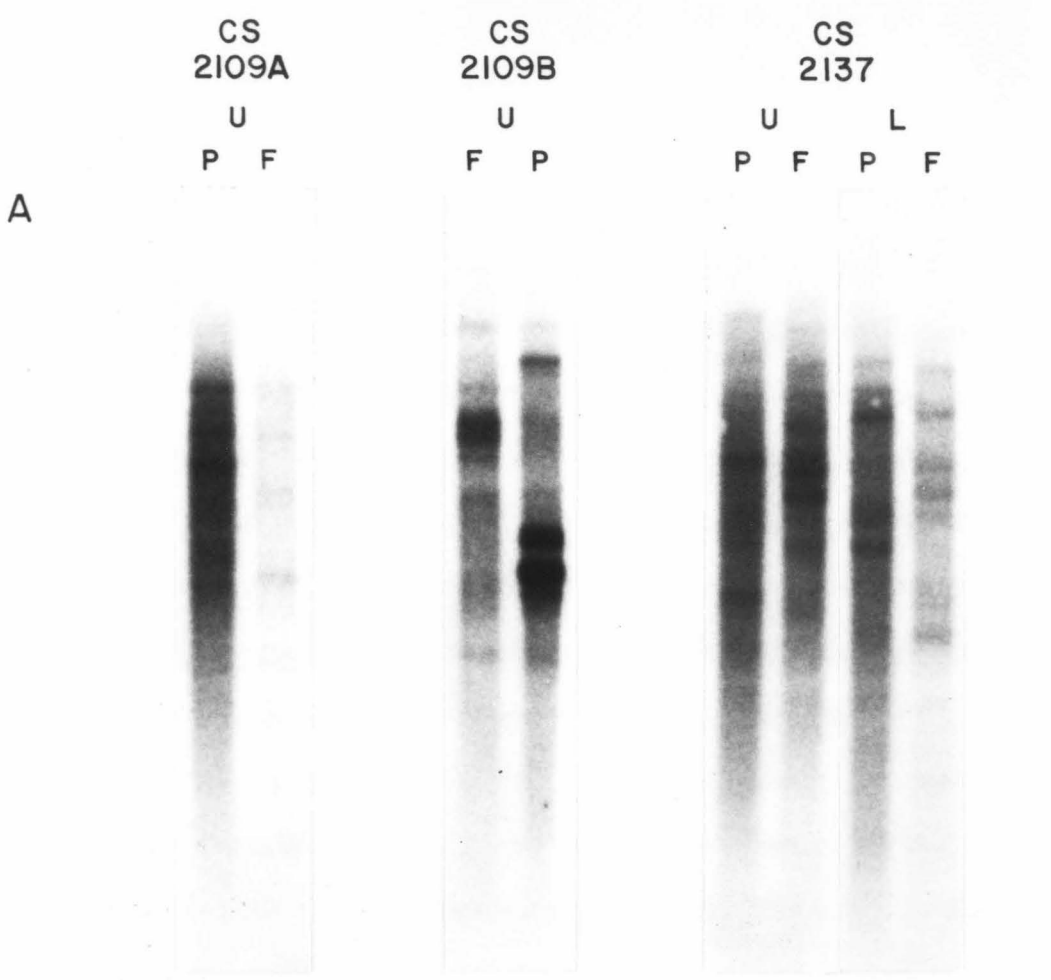

10HB
16PB
18AR1
5A2
E 16 B G P
E 16 B G
E 16 B G
E 16 B G

Figure 7

Gel blots of total poly(A)$^+$ RNAs isolated from different developmental stages of S. purpuratus, hybridized with restriction fragments of the cDNA clones (Figure 3). E, egg; 16, 16-cell; B, blastula; G, gastrula; P, pluteus. RNA (2 µg/lane) was denatured, electrophoresed, and transferred to nitrocellulose filters as described in Materials and Methods. Hybridization was carried out under 50% formamide criterion conditions (see Materials and Methods). The fragment used as a probe in each case is indicated (see Figure 3).


CS
2109A
U
P F
CS
2109B
U
F P
CS
2137
U L
P F P F
A

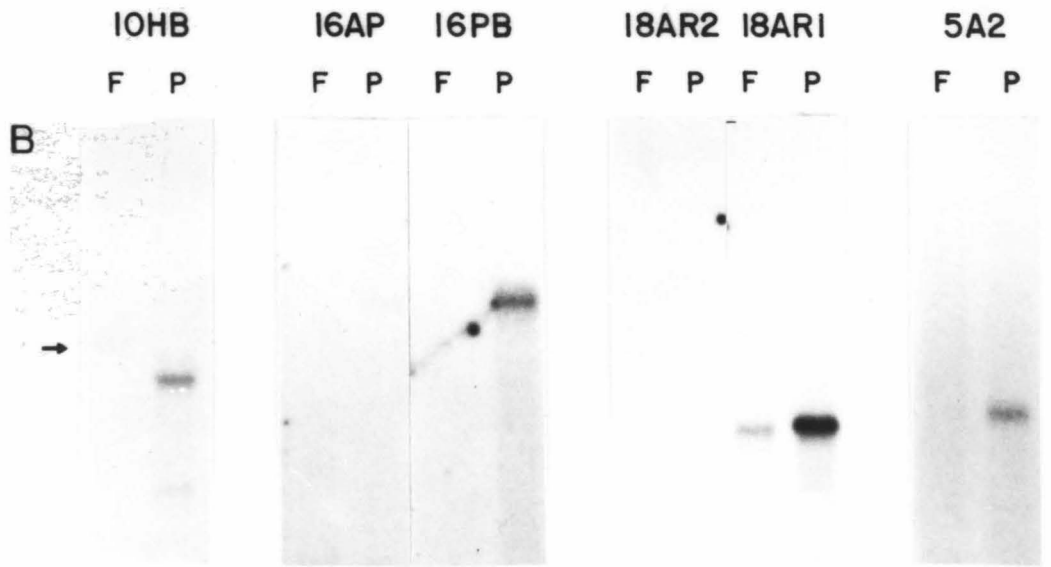
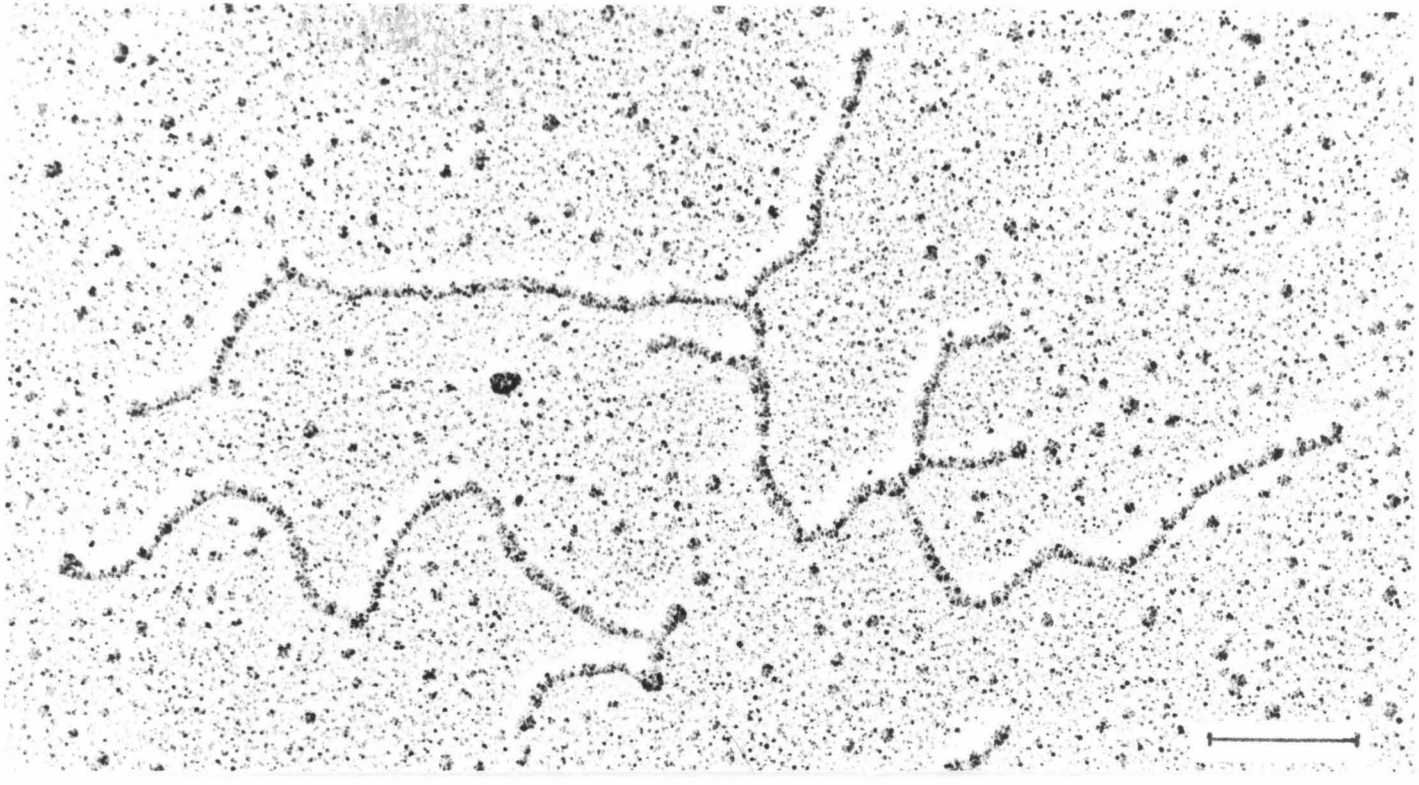

10HB
F P
16AP
F P
16PB
F P
18AR2 18AR1
F P F P
5A2
F P
B

Figure 8

Gel blots of egg poly(A)$^+$ RNA of S. purpuratus (P) and S. franciscanus (F), hybridized with separated strands of cloned repeat elements (A) or with selected restriction fragments (see Figure 3) from the cDNA clones (B). RNA (2 μg/lane) was denatured, electrophoresed, and transferred to nitrocellulose filters as described in Materials and Methods. Hybridization in A was carried out under 20% formamide criterion conditions, in B under 40% formamide conditions (see Materials and Methods). The DNA fragments used as probes in each case are indicated. In A, strand designations are as in Figure 1. In B, only the regions of the autoradiograms containing bands hybridizing to S. purpuratus RNA are shown; no additional hybridization to either RNA was observed outside these regions. Fragments 16AP and 16PB (p2109A-16) and 18AR1 and 18AR2 (p2109A-18) were radioactively labeled on opposite strands for these experiments (see Materials and Methods) as a test of the symmetry of representation in RNA of these single-copy sequences. The arrow to the left of the 10HB autoradiogram (B) indicates the position of a faint hybridizing band in S. franciscanus RNA.




5'
SC1
R1
SC2
R2
SC3
AAA···

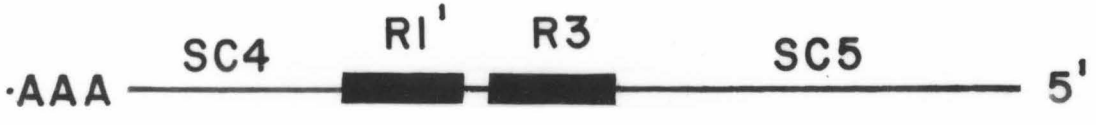

·AAA
SC4
R1'
R3
SC5
5'

Figure 9

Top: Electron micrograph of S. purpuratus egg poly(A)$^+$ RNA annealed to a Rot of approximately 1 M sec. Egg poly(A)$^+$ RNA was denatured, annealed, and spread for electron microscopy from 80% formamide as described by Costantini et al. (1980). The central structure appears to consist of two RNA molecules annealed to a third by complementary short repetitive sequences (see text). The bar indicates 1 kilobase of single-stranded RNA length, determined by measuring single-stranded ØX174 molecules included as internal standards and applying the corrections of Glass and Wertz (1980).

Bottom: Schematic drawings of two different interspersed polyadenylated RNA molecules, illustrating typical organizational features of such transcripts, as deduced from data presented in this chapter and in Costantini et al. (1980). The RNA strands are represented by solid lines; their 5' ends are indicated, as are their 3'-terminal poly(A) tracts. SC1, SC2, etc., represent different single-copy sequences forming much of the length of each transcript. R1, R2, and R3 represent repetitive sequences belonging to different families, interspersed within the transcripts. R1 and R1' represent complementary repeat sequences, transcribed from two different members of the same repeat family oriented oppositely in the genome with respect to the direction of transcription. The two molecules illustrated in the drawing could thus form an RNA-RNA duplex between R1 and R1', such as those shown in the electron micrograph (top).

Appendix

Cell, Vol. 21, 653–668, October 1980.

The Structure and Evolution of the Human β-Globin Gene Family

Argiris Efstratiadis
Department of Biological Chemistry
Harvard Medical School
Boston, Massachusetts 02115

James W. Posakony, Tom Maniatis, Richard M. Lawn* and Catherine O'Connell†
Division of Biology
California Institute of Technology
Pasadena, California 91125

Richard A. Spritz, Jon K. DeRiel,‡ Bernard G. Forget and Sherman M. Weissman
Departments of Genetics and Internal Medicine
Yale University School of Medicine
New Haven, Connecticut 06510

Jerry L. Slightom, Ann E. Blechl and Oliver Smithies
Laboratory of Genetics
University of Wisconsin
Madison, Wisconsin 53706

Francisco E. Baralle, Carol C. Shoulders and Nicholas J. Proudfoot§
MRC Laboratory of Molecular Biology
Hills Road
Cambridge CB2 2QH, England

Summary

We present the results of a detailed comparison of the primary structure of human β-like globin genes and their flanking sequences. Among the sequences located 5′ to these genes are two highly conserved regions which include the sequences ATA and CCAAT located 31 ± 1 and 77 ± 10 bp, respectively, 5′ to the mRNA capping site. Similar sequences are found in the corresponding locations in most other eucaryotic structural genes. Calculation of the divergence times of individual β-like globin gene pairs provides the first description of the evolutionary relationships within a gene family based entirely on direct nucleotide sequence comparisons. In addition, the evolutionary relationship of the embryonic ε-globin gene to the other human β-like globin genes is defined for the first time. Finally, we describe a model for the involvement of short direct repeat sequences in the generation of deletions in the noncoding and coding regions of β-like globin genes during evolution.

Present addresses:
* Genentech, Inc., 460 Pt. San Bruno Blvd., South San Francisco, California 94080.
† Frederick Cancer Center, Frederick, Maryland 21701.
‡ Fels Research Institute, Temple University School of Medicine, Philadelphia, Pennsylvania 19140.
§ Division of Biology, California Institute of Technology, Pasadena, California 91125.

Introduction

Comparisons of the amino acid sequences of a large number of globin polypeptides from different species indicate that the globin gene family evolved from a single ancestral gene by duplication and sequence divergence (Dayhoff, 1972). Concomitant with the evolution of functionally distinct globin genes, mechanisms for the regulation of their developmental expression must have been established. Thus a study of the molecular evolution of the human β-like globin genes and their flanking sequences may ultimately provide information relevant to the understanding of the evolution of regulatory pathways.

In the accompanying papers, we present the complete nucleotide sequences of the genes encoding embryonic ε-globin (Baralle, Shoulders and Proudfoot, 1980), fetal $^{G}\gamma$- and $^{A}\gamma$-globin (Slightom, Blechl and Smithies, 1980), and adult δ- (Spritz et al., 1980) and β-globin (Lawn et al., 1980). In this paper we compare the noncoding sequences within and surrounding each of the human β-like globin genes in order to identify regions of homology. The recognition of common sequence patterns within a family of related genes, in conjunction with functional in vitro (Weil et al., 1979; Manley et al., 1980) and in vivo (Hamer and Leder, 1979b; Mantei, Boll and Weissman, 1979; Wigler et al., 1979) assays for gene expression, may permit the identification of regulatory signals. To obtain an accurate description of the evolutionary relationships within the human β-globin gene family, we make use of an improved method for calculating the corrected percentage of divergence between coding sequences (Perler et al., 1980).

Discussion

Globin Gene Expression and Linkage

Several different α-like and β-like globin chains, which assemble in specific combinations to form functional hemoglobin tetramers, are synthesized at different times during human development. The embryonic ζ- (α-like) and ε- (β-like) globins which are found in embryos up to eight weeks of gestation are gradually replaced by the adult α chain and the fetal β-like chains $^{G}\gamma$ and $^{A}\gamma$ (for a recent review see Weatherall and Clegg, 1979). While synthesis of the α chain continues during fetal and adult life, the γ-globin polypeptides are replaced by the two adult β-like globins, β and δ, during the first six months after birth.

Genetic studies indicating the presence of two linked α-globin genes in most human populations (Bunn, Forget and Ranney, 1977; Weatherall and Clegg, 1979) were recently confirmed by blot hybridization (Orkin, 1978; Embury et al., 1979) and molecular cloning (Lauer, Shen and Maniatis, 1980). These genes have been assigned to chromosome 16 (Deis-

seroth et al., 1977). The linkage arrangement of the embryonic ζ- and the adult α-globin genes as well as an unexpressed α-globin pseudogene ($\psi\alpha1$) is shown in Figure 1 (Proudfoot and Maniatis, 1980; Lauer et al., 1980). All five α-like genes have the same transcriptional orientation and are arranged in the order $5'-\zeta_2-\zeta_1-\psi\alpha1-\alpha_2-\alpha_1-3'$.

A similar linkage arrangement has been established for the human β-like globin genes (Flavell et al., 1978; Lawn et al., 1978; Mears et al., 1978; Fritsch, Lawn and Maniatis, 1979, 1980; Bernards et al., 1979; Little et al., 1979a; Ramirez et al., 1979; Tuan et al., 1979) (Figure 1). These genes are located on chromosome 11 (Deisseroth et al., 1978; Jeffreys, Craig and Francke, 1979; Gusella et al., 1979; Lebo et al., 1979). In addition to the five β-like genes that correspond to known globin polypeptides, two other β-like sequences ($\psi\beta1$ and $\psi\beta2$) were detected by hybridization to cloned β- or γ-globin cDNA probes (Fritsch et al., 1980). Although the nature of these sequences has not yet been determined, it is possible that they are pseudogenes. As shown in Figure 1, the expressed β-like globin genes have the same transcriptional orientation and they are arranged together with $\psi\beta1$ and $\psi\beta2$ in the order $5'-(\psi\beta2)-\epsilon-{}^{G}\gamma-{}^{A}\gamma-(\psi\beta1)-\delta-\beta-3'$. (For a review of the chromosomal organization of the human α- and β-like globin genes, see Maniatis et al., 1980.)

A cluster of rabbit β-like globin genes has also been characterized and a similar overall organization is observed (Hardison et al., 1979; Lacy et al., 1979). This cluster consists of two embryonic genes ($\beta4$ and $\beta3$), a pseudogene ($\beta2$) and a single adult β-globin gene ($\beta1$), arranged in the order $5'-\beta_4-\beta_3-\beta_2-\beta_1-3'$ (Figure 1). Thus, although fetal genes are absent from the human α-like and the rabbit β-like clusters, all three gene families are similarly organized; the genes are arranged in the order of their developmental expression. In addition, in all cases a pseudogene is located in the intergenic region between the embryonic/fetal and adult globin genes (assuming that the human $\psi\beta1$ is a pseudogene). The same pattern of organization has recently been identified in the mouse β-globin gene family (Jahn et al., 1980).

The chromosomal arrangement of mammalian globin genes may be related to the mechanism of their differential expression (Bernards et al., 1979; Fritsch et al., 1979; Maniatis et al., 1980), or it may reflect their evolutionary history, or both.

Molecular Evolution of Human β-Like Globin Genes

Mutations have been estimated to occur in DNA sequences of individual organisms at a minimum rate of 7×10^{-9} per nucleotide site-year (Perler et al., 1980), but only a fraction of these changes are fixed in a population (nucleotide substitutions). When the coding regions of a pair of homologous genes are compared, codon to codon, two kinds of substitutions are evident—those resulting in amino acid replacements (replacement substitutions) and those leading to the appearance of synonymous codons (silent substitutions). The rate of accumulation of replacement (but not silent) substitutions has been studied extensively by comparing amino acid sequences within protein families. Sequence differences between homologous proteins are used to estimate the replacement site divergence of the corresponding genes by determining the minimum number of base changes necessary to generate the observed amino acid replacements. In addition, various mathematical correction methods have been devised to allow for multiple base change events within individual codons (for review see Wilson, Carlson and White, 1977; Dickerson and Geis, 1980). It is obvious that the rate of fixation of base changes cannot be estimated accurately from amino acid sequence comparisons due to the failure to detect certain replacement substitutions. An extreme example is the change of an AGC to a TCC codon, both encoding serine. Although no amino acid replacement is detected, two replacement substitutions have occurred in the DNA. Despite such limitations, amino acid sequence comparisons have been very useful in establishing molecular phylogenies and taxonomic relationships. Within a given protein family, replacement changes appear to accumulate linearly with divergence time as documented by the fossil record (the evolutionary clock hypothesis; Wilson et al., 1977).

A recent analysis at the DNA level, based on the divergence of insulin and globin gene sequences, indicates that replacement substitutions do in fact provide a reliable evolutionary clock (Perler et al., 1980). This analysis also suggests that silent substi-

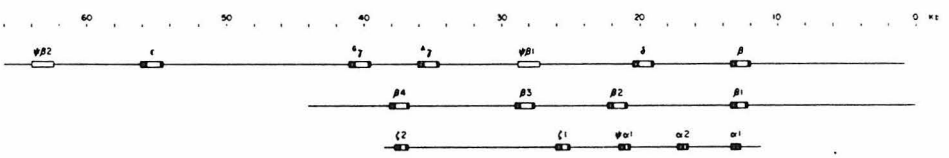


Figure 1. Linkage Maps of Three Mammalian Globin Gene Clusters

The locations of gene sequences are indicated by the solid (exons) and open (introns) boxes. The 5′ to 3′ direction of transcription is from left to right and the scale is in kb. From top to bottom are shown the human β-like, the rabbit β-like and the human α-like globin gene clusters (see text for references).

tutions can be used as an evolutionary clock, but only for recently diverged genes. We have used a clock based on replacement substitutions to construct an evolutionary tree of the human β-like globin genes.

To calibrate the evolutionary clock for globin genes, we calculated divergences at replacement substitution sites for pairs of genes for which the evolutionary divergence times are known from the fossil record. Silent site divergences were also calculated. The values were then corrected for multiple base change events, assuming that these are Poisson-distributed (Holmquist, 1972; Kimura and Ohta, 1972; Salser, 1978; see Perler et al., 1980, for a detailed description of the methodology). These corrected percent divergences are shown in Table 1. The following is a detailed description of the replacement site calibration points.

—The average divergence between pairs of α- or β-globin genes of human, rabbit and mouse (10%) was assumed to correspond to the mammalian radiation, which occurred approximately 85 million years (MY) ago (Romero-Herrera et al., 1973).

—The average divergence between mammalian and chicken α- and β-globin genes (23%) was assumed to correspond to 270 MY. Data from the fossil record place the bird-mammal divergence at 250–300 MY ago (Dickerson, 1971; Moore et al., 1976; Wilson et al., 1977).

—The divergence between the α- and β-globin sequences within a species was assumed to correspond to 500 MY. Since the cyclostomes (lamprey and hagfish) have a single globin chain, while the carp has both α and β chains, it is assumed that the α/β duplication occurred some time during the evolution of bony fish, approximately 500 MY ago (Dickerson and Geis, 1980). The average intraspecies α/β diver-

Table 1. Corrected Percent Divergences of Coding Sequences of Mammalian and Chicken Globin Genes

Gene Pair	Replacement Sites	Silent Sites
α-Globin		
Human/rabbit	11	32
Human/mouse	8.4	83
Rabbit/mouse	11	82
Human/chicken	21	75
Rabbit/chicken	23	64
Mouse/chicken	20	87
β-Globin		
Human/rabbit	5.6	46
Human/mouse	13	49
Rabbit/mouse	12	64
Human/chicken	23	70
Rabbit/chicken	23	81
Mouse/chicken	27	79
Rabbit α/β	49	92
Mouse α/β	51	120
Chicken α/β	51	87
Human α/β	46	90
α/δ	47	100
$\alpha/^{G}\gamma$	51	103
α/ε	51	91
Human $^{G}\gamma/^{A}\gamma$	0.35 (0, 0, 1.2)	0 (0, 0, 0)
β/δ	3.7 (5.5, 2.2, 5.1)	32 (16, 16, 112)
$^{G}\gamma/\epsilon$	10 (16, 6.1, 14)	61 (58, 43, 191)
$\beta/^{G}\gamma$	18 (28, 15, 20)	74 (87, 54, 145)
$\delta/^{G}\gamma$	19 (31, 14, 21)	75 (93, 53, 190)
β/ε	16 (15, 16, 16)	61 (116, 66, 66)
δ/ε	17 (18, 17, 18)	62 (137, 59, 82)
Mouse β^{maj}/β^{min}	3.1 (4.9, 3.7, 2.3)	11 (0, 17, 15)

The corrected percent divergence of replacement and silent substitution sites in each pair of coding sequences was calculated by Perler et al. (1980). These calculations include corrections for multiple base change events at single sites (Holmquist, 1972; Kimura and Ohta, 1972); consequently, some of the values in the table exceed 100%. In the comparison of an α gene versus a β-like gene, we excluded from the calculation the codons of one sequence corresponding to gaps in the other (for the alignment see Dayhoff, 1972). For the comparisons of the human β-like globin genes and also mouse β^{maj}/β^{min}, the table shows first the divergence calculated for the total coding sequence, followed by the divergences of the three exons separately (numbers in parentheses). Except for the mouse β^{maj}/β^{min} gene comparison, the mouse β^{maj} gene sequence was used in all pairwise comparisons listed as involving the mouse β-globin gene. Sequence data for non-human genes or cDNAs are from Hardison et al. (1979) (rabbit β), Konkel et al. (1979) (mouse β), Richards et al. (1979) (chicken β), Heindell et al. (1978) (rabbit α) and Nishioka and Leder (1979) (mouse α). The chicken α-globin sequence (Salser et al., 1979) represents a transcribable α-globin gene, but it does not correspond in its entirety to the two known amino acid sequences for chicken α-globins. The human α-globin cDNA sequence is from Forget et al. (1979).

Cell
656

gence for rabbit, mouse and chicken is 50%; the average divergence of the human α/β, α/δ, α/γ and α/ϵ gene pairs is 49%.

These replacement site calibration points are plotted in Figure 2 (solid circles), along with the corresponding points for divergence at silent sites (open circles). A straight line (R) can be drawn through the origin and the replacement site points. The line serves as a molecular clock. Its slope gives the rate of fixation of replacement substitutions in globin genes, and corresponds to a unit evolutionary period (UEP) of approximately 10. The UEP is the time in MY required for the fixation of a 1% divergence between two initially identical sequences (Wilson et al., 1977).

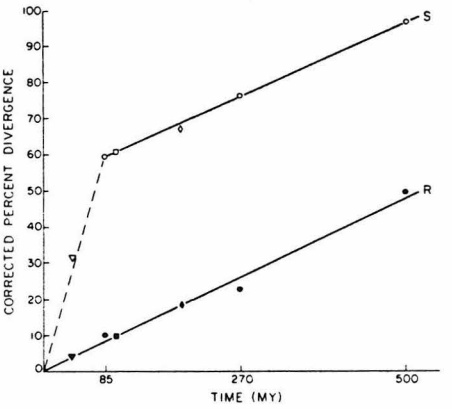


Figure 2. Calibration Curves for Divergence of Globin Coding Sequences

The corrected percent divergences at silent (open symbols) and replacement (closed symbols) substitution sites (Table 1) are plotted against divergence time in millions of years (MY). The circles at 85, 270 and 500 MY correspond respectively to divergences calculated from the following comparisons: interspecies comparisons of α- and β-globin genes of different mammals; comparisons of α- and β-globin genes of mammals with those of chicken; intraspecies comparisons of α-globin genes with β-globin genes. The time values are derived from generally accepted interpretations of the fossil record (see text for references).

The solid lines (S and R) were drawn through the calibration points without fitting. Line R was made to pass through the origin. Its slope gives the rate of fixation of replacement substitutions in globin genes, and corresponds to a unit evolutionary period (UEP) of approximately 10. [The UEP is the time in MY required for two initially identical sequences to acquire a 1% sequence divergence (Wilson et al., 1977).] It should be noted that alternative interpretations of the fossil data would allow a more linear relationship among the replacement site calibration points. For example, certain paleontological lines of evidence indicate that the mammalian radiation can be positioned at 100 MY (McKenna, 1969) and the bird/mammal divergence between 200 and 250 MY (Bakker, 1975; Dickerson and Geis, 1980). Since line S does not extrapolate to the origin, we presume that the silent sites diverge linearly with a high rate for approximately 100 MY (dashed line). The significance of the silent site calibration was examined as follows. The replacement site divergences for the β/δ (triangles), γ/ϵ (squares) and β or δ versus γ or ϵ (diamonds) gene pairs were first placed on line R. The corresponding silent site divergences (matching open symbols) were then plotted directly above. Note that the latter points fall close to the silent site calibration lines.

Using this estimate of the UEP, we have constructed an evolutionary tree of the human β-like globin genes. We calculated replacement and silent site divergences between the human genes as described above for the evolutionary clock calibration points. The results of these calculations are shown in Table 1. When the corrected percent divergence at replacement sites of a given pair of genes is multiplied by the UEP, the result is an estimate of the time since the two genes began to diverge. Our analysis indicates that the β/δ divergence occurred approximately 40 MY ago. This value is significantly greater than previous estimates (for example see Dickerson and Geis, 1980). The γ/ϵ divergence is older (100 MY ago), and within the limits of our calculations, we consider that the divergence of the human fetal and embryonic globin genes occurred at approximately the onset of the mammalian radiation. The β and δ genes are nearly equally divergent from the γ and ϵ genes (Table 1). This divergence occurred approximately 200 MY ago, some time after the bird-mammal divergence, during the evolution of the reptiles which gave rise to mammals (Dickerson and Geis, 1980).

Figure 3 presents an evolutionary tree constructed from the divergence times of individual gene pairs

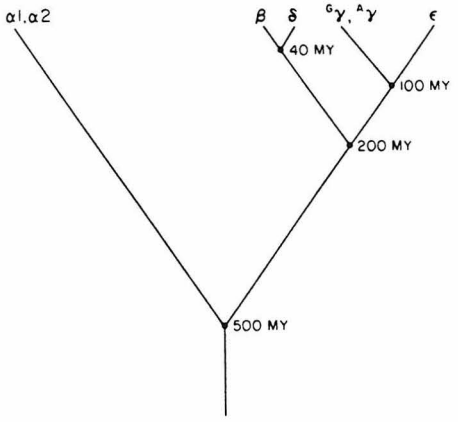


Figure 3. An Evolutionary Tree for Human β-Like Globin Genes

Corrected percent divergences at replacement sites of individual gene pairs (Table 1) were multiplied by our estimate of the UEP for globin genes to yield the corresponding divergence times (see text). An evolutionary tree was then constructed by a simultaneous equation method (Dickerson and Geis, 1980). Distances along the branches are additive and proportional to evolutionary time. Branch points represent the times in millions of years (MY) from the present at which genes or gene lineages began to diverge. As discussed in the text, these branch points do not necessarily represent gene duplication events.

according to the simultaneous equation method described by Dickerson and Geis (1980). The tree represents the first description of the evolutionary relationships within a gene family based entirely on direct nucleotide sequence comparisons. In addition, the evolutionary relationship of the ε gene to the other human β-like globin genes is defined for the first time.

The branch points of the tree (Figure 3) may represent times at which gene duplication events took place. However, it is possible that the initial products of a given duplication were corrected against each other for an unknown period of time. In such a case, the branch point would correspond to the time since the last gene correction event. Evidence for the existence of a gene correction mechanism is provided by structural analysis of the human fetal globin genes, $^{G}\gamma$ and $^{A}\gamma$ (Jeffreys, 1979; Little et al., 1979b; Slightom et al., 1980). A study of the complete nucleotide sequences of these two genes has led Slightom et al. (1980) to propose a specific model for intrachromosomal gene conversion. In this case the time of the initial gene duplication event is not the same as the time of the last intergenic exchange. C.-K. J. Shen, J. L. Slightom and O. Smithies (unpublished results) have shown that the $^{G}\gamma$ and $^{A}\gamma$ genes are part of a large duplication of about 5 kb, and they have found that over much of the duplication region the DNA sequences of the duplicates have diverged by 10–20%. Yet within the 5 kb duplication there are approximately 1.5 kb of virtually identical DNA sequence which appear to have been involved in a recent intergenic exchange event. This sequence includes the 5′ two thirds of the genes (Slightom et al., 1980). Outside the 5 kb duplication unit, the sequences are approximately 75% divergent. These observations indicate that the gene correction unit is not identical to the gene duplication unit.

The evolutionary tree of human β-like globin genes (Figure 3) indicates that distinct embryonic and fetal β-like sequences appeared 100 MY ago, at approximately the time of the mammalian radiation. It is possible that the appearance of two different β-like genes specialized to function in the embryo or the fetus, respectively, was related to the novel physiological requirements of placental development (Bunn et al., 1977). The fact that some mammals such as rabbits and mice do not appear to have genes which are expressed specifically in the fetus may be due to the relatively short gestation times and relative immaturity of the neonate in these species.

The tree also shows that the adult β and δ genes diverged from each other approximately 40 MY ago (Figure 3). Amino acid sequence (Dayhoff, 1972) and blot hybridization (S. L. Martin, personal communication) data indicate that δ-globin genes are present in all higher primates. This suggests that the β/δ divergence occurred prior to the divergence of the phylogenetic lines leading respectively to the New World monkeys and to the Old World monkeys, great apes and humans. This latter divergence has been placed at 35 to 40 MY ago (D. Pilbeam, personal communication), a time which is virtually identical to that calculated here for the β/δ gene divergence.

Comparisons of Individual Exons of β-Like Globin Genes

Individual globin gene exons are thought to encode functional protein domains (Blake, 1979; Gilbert, 1979; Eaton, 1980). In fact, Craik, Buchman and Beychok (1980) demonstrated that the polypeptide encoded by the middle exon of human α- or β-globin genes specifically binds heme in vitro. It is therefore possible that individual exon sequences diverge at different rates during evolution. To examine this possibility, we calculated divergences from pairwise comparisons of individual exons of β-like globin genes (Table 1). This analysis revealed that within four of the gene pairs examined (β/δ, γ/ε, β/γ and δ/γ) the middle exon exhibits less replacement site divergence than the outer two exons. However, the same is not true in the case of the β/ε, δ/ε or mouse β^{maj}/β^{min} gene pairs. Thus there does not appear to be a consistent pattern of differential divergence of individual globin gene exons. .

Comparison of the Flanking, Noncoding and Intervening Sequences of β-Like Globin Genes

In the following sections we discuss detailed comparisons of the noncoding sequences within and surrounding mammalian β-like globin genes. The purpose of these comparisons is to identify conserved sequences which might have functional significance, and to examine the sequence divergence of the noncoding regions in relation to that of the coding regions (see Tables 1 and 2). The comparisons were made after establishing the best alignment of the sequences, in pairs or as a group, either directly (see legend to Table 2) or by computer methodology (see legend to Figure 7). Analysis of a large group of related DNA sequences in this manner reveals predominant patterns of homology which would not be evident from examination of a single pair of genes.

5′ Flanking Regions

An alignment of the sequences located 5′ to the mRNA capping site of ten different mammalian β-like globin genes is presented in Figure 4. The precise location of the capping site is known only for the adult mouse, human and rabbit β-globin genes (for references see Baralle and Brownlee, 1978) and for the human fetal γ-globin genes (Chang et al., 1978). The corresponding site in the other β-like globin genes was assigned by analogy.

The first region of strong homology in the 5′ direction from the capping site is an AT-rich sequence, containing the trinucleotide ATA, which is flanked by

Table 2. Corrected Percent Divergences of Noncoding Sequences of Mammalian β-Like Globin Genes

Gene Pair	5′ Flanking Region	5′ Noncoding Region	IVS1	IVS2	3′ Noncoding Region
Human $^{G}\gamma/^{A}\gamma$	0	1.9	0	1.4 (4.5)[a]	7.0
β/δ	34	8.5	12		59
$^{G}\gamma/\epsilon$	43	65	46		71
$\beta/^{G}\gamma$	45	49			68
$\delta/^{G}\gamma$	63	57			86
β/ϵ	56	35	78		57
δ/ϵ	72	49			65
Mouse β^{maj}/β^{min}	36	20	2.6	16	36

[a] The number in parentheses is the percent divergence calculated over just that part of IVS2 in which base substitutions occur (see Slightom et al., 1980).

Sequence divergences for the noncoding and flanking regions were calculated from the alignments shown in Figures 4, 5, 6 and 9. (IVS1 and IVS2 divergences for the mouse β^{maj}/β^{min} gene pair were calculated from the alignment of Konkel et al., 1979.) Direct alignment (for Figures 4, 5 and 9) was carried out as follows. Sequences were first aligned in pairs. Gaps were introduced in one sequence relative to the other where necessary to accommodate length differences and to maintain alignment of homologous segments on either side of the gaps. Alternative alignments of a given sequence segment were examined to identify the most likely homolog of that segment and to minimize the number of gaps introduced. Pairwise alignments were then reconciled into a single group alignment by identifying the predominant patterns of homology of the group as a whole. The alignment of each sequence was adjusted where necessary to conform to these patterns. Computer-assisted alignment (for Figure 6) was carried out as described in the legend to Figure 7. In calculating divergences, gaps (or insertions) in one sequence relative to the other (introduced during alignment; see above) were ignored. Thus the numbers in the table refer to sequence divergence resulting from base substitution. A correction factor for multiple base change events at single sites was applied to the raw divergence values to give the percent divergences shown in the table.

GC-rich segments (Figure 4). This sequence, which we will refer to as the ATA box, was first recognized in the Drosophila histone genes (D. Hogness, personal communication; Goldberg, 1979) and was subsequently identified in many other eucaryotic genes (see Figure 4). The first A of the ATA trinucleotide is located approximately 30 nucleotides upstream from the capping site. The conserved location of the ATA box and its similarity to a sequence characteristic of all procaryotic promoters (Pribnow, 1979) led many investigators to speculate that it might have a role in the initiation of transcription in eucaryotes. However, the absence of the ATA box 5′ to the regions encoding certain viral mRNAs suggests that the ATA box is not essential for transcription in these cases (Baker et al., 1979). The absence of the ATA box has been correlated with microheterogeneity of the capped mRNA 5′ end (Baker et al., 1979). This correlation is consistent with the observation that the deletion of the ATA box from a cloned sea urchin H2A histone gene does not abolish transcription of the gene in Xenopus oocytes, but leads to the production of transcripts with heterogeneous 5′ termini (Grosschedl and Birnstiel, 1980). On the other hand, Wasylyk et al. (1980) have recently shown that the ATA box is required for the transcription of the chicken conalbumin gene in an in vitro system containing calf thymus RNA polymerase II and a cell-free extract prepared from HeLa cells. These contrasting results may be due to the use of different transcription assays (in vivo versus in vitro) or to actual differences in the promoters studied. In any case, the functional significance suggested by the conserved sequence and location of the ATA box and by the results of transcription studies remains to be established definitively.

In the 5′ direction from the ATA box, there is a second region of strong homology which includes the sequence CCAAT ("CCAAT box"). This sequence is present in all the β-like globin genes of Figure 4 except δ (which contains the sequence CCAAC), and it also occurs at a corresponding position in the mouse (Nishioka and Leder, 1979) and human (Liebhaber et al., 1980) α-globin genes. A 16 nucleotide region containing the CCAAT box is duplicated in the γ-globin genes (Figure 4).

As mentioned by Benoist et al. (1980), a sequence similar to the globin CCAAT box is found 5′ to the capping site in several other eucaryotic genes. An updated list of such sequences is shown in Figure 4. As in the case of the ATA box, the widespread occurrence of the CCAAT box and its location with respect to the capping site suggest that it may play a role in transcription initiation. However, deletion of a region containing a CAAT sequence 5′ to the sea urchin H2A histone gene does not abolish, and may even increase, transcription of this gene in the Xenopus oocyte system (Grosschedl and Birnstiel, 1980). Similarly, deletion of the corresponding sequence 5′ to the chicken conalbumin gene (Wasylyk et al., 1980) and the human β-globin gene (V. Parker, N. Proudfoot, M. Shander and T. Maniatis, unpublished results) does not prevent specific in vitro transcription.

Alignment of the sequences in the region 5′ to the ATA box reveals considerable homology among all the mammalian β-like globin genes shown in Figure 4. The best alignment was obtained by assuming that the γ- and ε-globin genes contain an additional 11 and 8 nucleotides, respectively, immediately 5′ to the ATA box. It was also necessary to assume that the δ gene lacks 7 nucleotides adjacent to the ATA box.

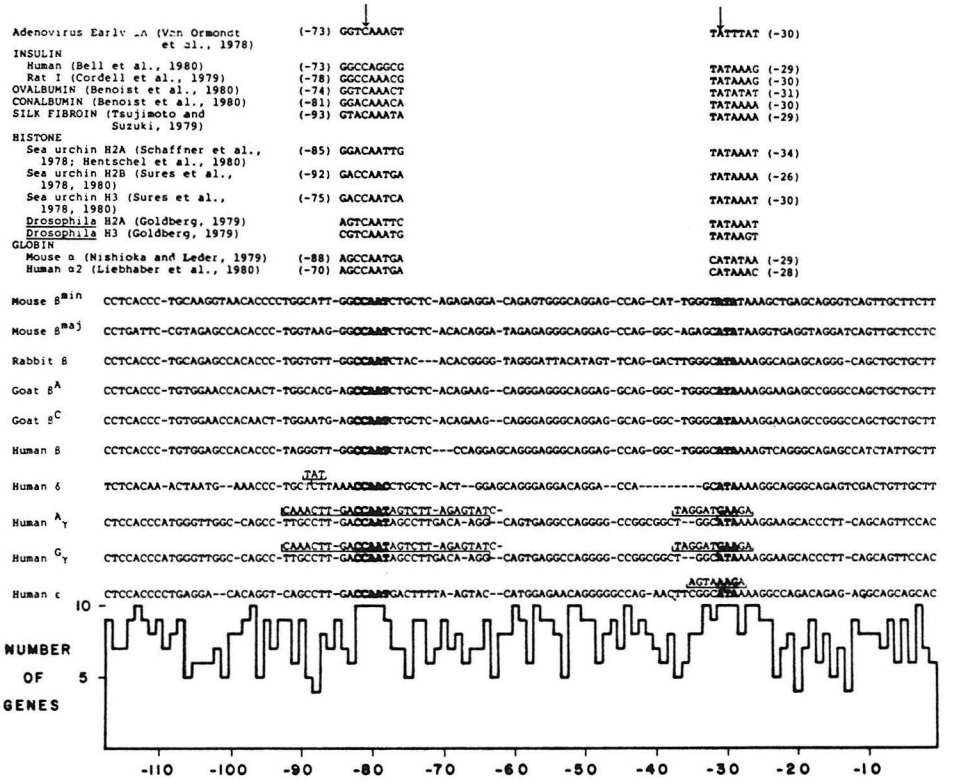


Figure 4. Alignment of the 5′ Flanking Sequences of Mammalian β-Like Globin Genes

An alignment of the sequences preceding the mRNA capping site of ten different mammalian β-like globin genes is shown. Dashes indicate gaps introduced to achieve the most consistent alignment (see legend to Table 2). Similarly, the nucleotides above a given sequence were displaced to give the best alignment of the sequences on either side. Regions of sequence homology which are common to all globin genes sequenced to date (the ATA and CCAAT boxes) are indicated by shading. Similar sequences found in the 5′ flanking regions of other eucaryotic genes are shown above the alignment. Some of the sequences homologous to the CCAAT box are different from those pointed out by Benoist et al. (1980). Numbers in parentheses give the position of the indicated bases (arrows) in nucleotides from the mRNA capping site (where known). A histogram is shown below the alignment to illustrate the extent of homology at each nucleotide position. The ordinate is frequency of occurrence of the predominant nucleotide at each position. References for the β-like globin sequences are presented in the legend to Table 1. The goat sequences are from Haynes, Rosteck and Lingrel (1980).

Thus considerable heterogeneity in the distance between the ATA and CCAAT boxes is observed among the human β-like globin genes, as well as among other eucaryotic genes (Figure 4).

The sequences upstream from the CCAAT box are remarkably similar among the adult β-globin genes of different mammals. This conserved sequence appears to be specific to these genes, since it is not shared with the human embryonic/fetal β-like globin genes, the human or mouse α-globin genes or other eucaryotic genes (see Figure 4).

Corrected percent divergences of the 5′ flanking regions are shown in Table 2 for pairs of human β-like globin genes and for the mouse β^{maj}/β^{min} gene pair.

The 5′ flanking sequences of the δ-globin gene show significant deviations from the sequence features common to the other β-like globin genes. The amount of δ-globin mRNA precursor in nuclei from unfractionated bone marrow cells is 10 fold less than the amount of β-globin precursor (Kantor, Turner and Nienhuis, 1980). This could be due to differences in the rates of either (or both) transcription or turnover of δ- and β-globin RNA. Perhaps the dissimilarities in the 5′ flanking sequence of the δ- and β-globin genes are related to their respective levels of expression.

5′ Noncoding Region

Figure 5 shows an alignment of the DNA sequences corresponding to the 5′ noncoding regions of mammalian globin mRNAs. The number of bases between

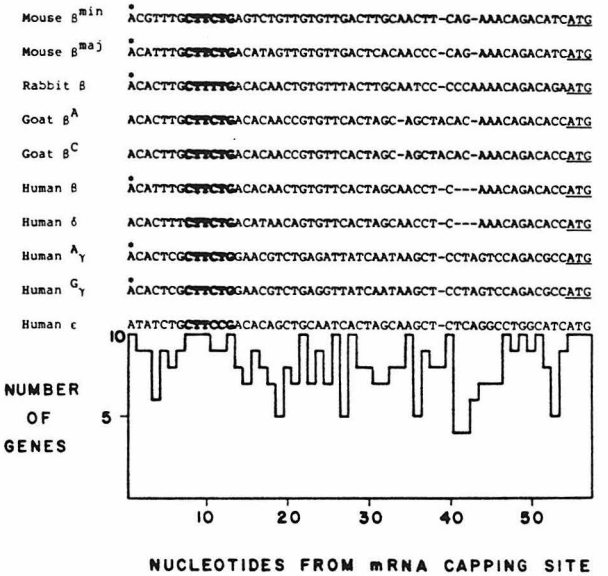


Figure 5. Alignment of the 5′ Noncoding Sequences of Mammalian β-Like Globin Genes

An alignment of the sequences between the mRNA capping site and the initiator ATG (underlined) of ten different mammalian β-like globin genes is shown. Capping sites (where known) are indicated by asterisks. Dashes indicate gaps introduced as required for alignment (see legend to Table 2). The location of the sequence CUUPyUG common to the 5′ noncoding regions of most eucaryotic mRNAs (Baralle and Brownlee, 1978) is indicated by shading. A histogram is shown below the alignment to illustrate the extent of homology at each nucleotide position. The ordinate is the frequency of occurrence of the predominant nucleotide at each position.

the capping site and the initiator ATG is relatively constant (50–53 bp) among the β-like globin genes. The sequence CUUPyUG, common to the 5′ noncoding regions of most eucaryotic mRNAs (Baralle and Brownlee, 1978), is found in the β-like globin genes seven nucleotides 3′ to the capping site. None of the genes contains an ATG triplet between the capping site and the initiation codon, consistent with the proposal (Kozak, 1978) that translation is initiated with the ATG closest to the 5′ end of the mRNA.

Corrected percent divergence values for the 5′ noncoding region are shown in Table 2.

Intervening Sequences

All β-like globin genes studied thus far contain a small and a large intervening sequence (Jeffreys and Flavell, 1977; Konkel, Tilghman and Leder, 1978; Lawn et al., 1978; Smithies et al., 1978; Tilghman et al., 1978; Proudfoot and Baralle, 1979; Konkel, Maizel and Leder, 1979; Hardison et al., 1979; van Ooyen et al., 1979; and the accompanying papers). In all cases the small intron (IVS1) is located between the codons for amino acids 30 and 31, while the large intron (IVS2) is found between codons 104 and 105. Both introns are variable in size. IVS1 ranges from 116 bp in the mouse β genes to 130 bp in the human β gene, and IVS2 from 573 bp in the rabbit β gene to 904 bp in the human $^{G}\gamma$ gene (Table 3). Among the human β-like genes, the size of IVS2 varies over a much narrower range, from 854 bp (ε gene) to 904 bp ($^{G}\gamma$ gene).

Table 3. Sizes of Intervening Sequences in Mammalian β-Like Globin Genes

Gene	Size (bp) IVS1	IVS2
Human β	130	850
δ	128	889
$^{A}\gamma$	122	866, 876*
$^{G}\gamma$	122	886, 904*
ε	122	850
Rabbit β	126	573
Mouse β^{maj}	116	653
β^{min}	116	628

* See Slightom et al., 1980.

Intervening sequences were defined by the rule of Breathnach et al. (1978), with the dinucleotide GT at their 5′ end and the dinucleotide AG at their 3′ end. Sources for the non-human sequence data are given in the legend to Table 1.

To compare the intron sequences and to estimate divergences, we attempted to align the introns of each pairwise combination of the human β-like genes. We made use of a computer program which prints out a matrix of dots indicating segments of homology between two sequences (R. J. Britten, unpublished). A similar program has been described by Konkel et al. (1979). The small introns of the $^{G}\gamma$- and $^{A}\gamma$- globin genes on one chromosome are identical (Figure 6). In the case of the other human genes, IVS1 alignments were possible only for the β/δ, γ/ε and β/ε gene pairs (Figure 6). For example, in the dot matrix shown in

Figure 6. Pairwise Alignments of the Small Intervening Sequences of Human β-Like Globin Genes

Alignments of the small intervening sequences of the $^{G}\gamma/^{A}\gamma$, β/δ, γ/ϵ and β/ϵ gene pairs are shown. Vertical lines are drawn between homologous bases of each pair. Dashes indicate gaps introduced as required for alignment (see legend to Figure 7). The 5′ and 3′ ends of the introns are defined by the dinucleotides GT and AG, respectively (Breathnach et al., 1978).

Figure 7A, the best alignment of IVS1 for the β/δ gene pair (including the 2 bp insertion/deletion) is unmistakable, while no meaningful alignment of IVS1 for the β/γ gene pair is evident from the pattern shown in Figure 7B. The corrected percent divergences of IVS1 of the pairs which could be aligned are shown in Table 2. For comparison, we also calculated the IVS1 divergence of the mouse β^{maj}/β^{min} gene pair. Alignments of IVS2 are possible only for the human $^{G}\gamma/^{A}\gamma$ and mouse β^{maj}/β^{min} gene pairs (see Slightom et al., 1980, and Konkel et al., 1979).

Our analysis indicates that in recently diverged pairs of genes the small intron shows much less sequence divergence than the large intron. We can make this statement quantitatively for the human $^{G}\gamma/^{A}\gamma$ and mouse β^{maj}/β^{min} comparisons (Table 2). In the latter case, the divergences of the small and large introns differ by more than 6 fold (2.6% versus 16%). The above conclusion is also justified for the human β and δ genes. The small intron divergence is quite low (12%), while on the basis of a dot matrix analysis (Figure 7C), the large intron cannot be aligned meaningfully, indicating a much higher divergence. Finally, it is likely that there is a significant difference between the small and large intron divergences of the γ/ϵ gene pair, since one intron could be aligned readily and the other could not.

How can these divergence differences be explained? One possibility is that, as in the case of the human γ-globin gene pair, 5′ regions of the human β/δ, human γ/ϵ and mouse β^{maj}/β^{min} gene pairs were corrected against each other for some time during their evolutionary history. This hypothesis predicts that the divergence at silent substitution sites in the 5′ coding regions of these gene pairs will be significantly less than that of the 3′ uncorrected coding regions. In addition, the percent divergence of the small intron should be similar to the silent site divergence of the corrected exons. The latter prediction is based on the observation that for recently diverged gene pairs the silent site divergence and the noncoding sequence divergences are similar (Perler et al., 1980). The data presented in Table 4 are consistent with these predictions. The silent site divergence in the first two exons of the human β/δ gene pair is 16%, and the divergence of IVS1 is 12%. In contrast, IVS2 and the silent sites of exon 3 are much more divergent. Thus, if a region-specific correction mechanism did act on the β/δ gene pair, the boundary between corrected and uncorrected sequences would have been located somewhere near the junction of exon 2 and IVS2. A similar pattern of divergences is observed in the human γ/ϵ gene pair (Table 4). In the case of the mouse β^{maj}/β^{min} gene pair, exon 1 and IVS1 show very little divergence, while exon 2, IVS2 and exon 3 are more divergent. Thus for this gene pair the putative boundary between corrected and uncorrected regions would be near the junction of IVS1 and exon 2. When a similar analysis of the human $^{G}\gamma/^{A}\gamma$ gene pair is made, the silent sites and intron sequences located 5′ to a site within the large intron are identical, while the intron sequences located 3′ to this site are 4.5% divergent. However, no silent changes are found in exon 3.

An alternative explanation for the differential divergences of IVS1 and IVS2 of the gene pairs shown in Table 4 is that the two introns are subject to different degrees of selection. For example, it is possible that the structure of the 5′ region of globin nuclear mRNA precursors has a more important role in processing than does the 3′ region.

Both the size variability (Table 3) and the sequence alignments (Figure 6) of the introns indicate that they are evolving by a combination of base substitution and deletion/insertion as suggested by Konkel et al. (1979) and van Ooyen et al. (1979). This undoubtedly contributes to our inability to align the large introns of the human β-like globin genes (see above). In a later section, we will present a model for the generation of deletions in introns and other parts of the gene.

The junctions between exons and introns in the globin genes are regions of particular interest, since they are the sites of action of the processing mechanism which removes intron sequences from the nu-

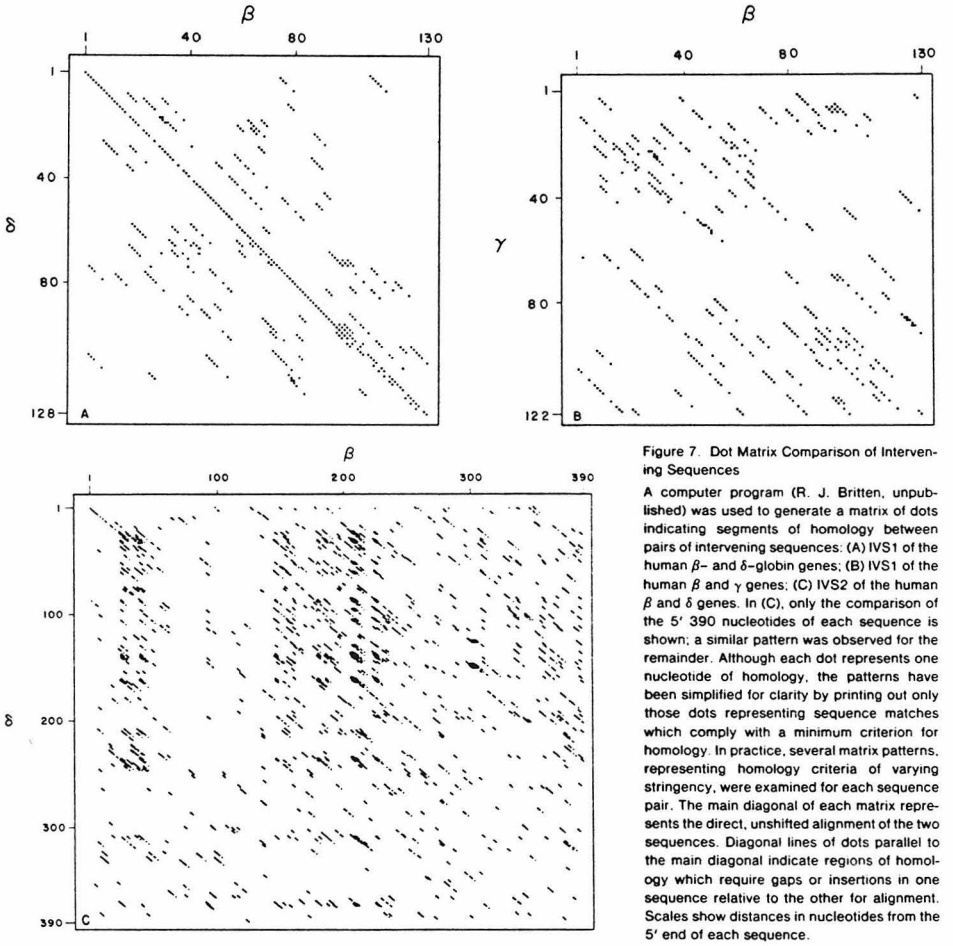


Figure 7. Dot Matrix Comparison of Intervening Sequences

A computer program (R. J. Britten, unpublished) was used to generate a matrix of dots indicating segments of homology between pairs of intervening sequences: (A) IVS1 of the human β- and δ-globin genes; (B) IVS1 of the human β and γ genes; (C) IVS2 of the human β and δ genes. In (C), only the comparison of the 5′ 390 nucleotides of each sequence is shown; a similar pattern was observed for the remainder. Although each dot represents one nucleotide of homology, the patterns have been simplified for clarity by printing out only those dots representing sequence matches which comply with a minimum criterion for homology. In practice, several matrix patterns, representing homology criteria of varying stringency, were examined for each sequence pair. The main diagonal of each matrix represents the direct, unshifted alignment of the two sequences. Diagonal lines of dots parallel to the main diagonal indicate regions of homology which require gaps or insertions in one sequence relative to the other for alignment. Scales show distances in nucleotides from the 5′ end of each sequence.

clear precursors during the production of mature mRNA. Figure 8 shows an alignment of the exon/intron boundary regions of the human β-like globin genes, along with those of the rabbit and mouse β-globin genes. In agreement with the rule proposed by Breathnach et al. (1978), the 5′ ends of both introns in all the genes contain the dinucleotide GT, and the 3′ ends contain the dinucleotide AG. The globin gene exon/intron junction sequences also show good agreement with consensus sequences from other eucaryotic genes (Breathnach et al., 1978; Seif, Khoury and Dhar, 1979; Lerner et al., 1980). The 5′ boundary of IVS2 of the β-like globin genes is unusual in the preservation of a 7 bp sequence of which 6 bp are within the intron (as defined by the GT dinucleotide).

Hamer and Leder (1979a) have shown that at least one splicing event is necessary for the production of a functional globin mRNA in vivo. The lack of conservation of the majority of IVS2 sequences among most mammalian β-like globin genes suggests that only the exon/intron junction sequences and some unknown

Table 4. Corrected Percent Divergences in Silent Sites and Intervening Sequences

Gene Pair	Exon 1	IVS1	Exon 2	IVS2	Exon 3
Mouse β^{maj}/β^{min}	0	2.6	17	16	15
Human β/δ	16	12	16	—[a]	112
Human γ/ϵ	58	46	43	—[a]	191

[a] The inability to align these sequences indicates a large divergence. These data are taken from Tables 1 and 2.

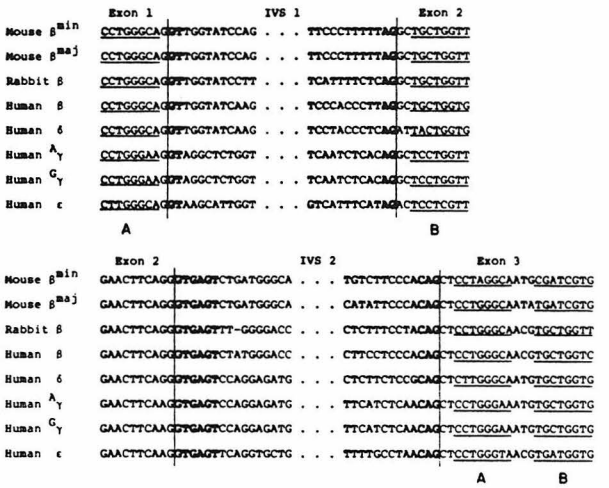


Figure 8. Alignment of Sequences surrounding Intron/Exon Junctions of Mammalian β-Like Globin Genes

An alignment of the sequences surrounding the four intron/exon junctions of eight mammalian β-like globin genes is shown. Vertical lines indicate the intron/exon junctions, following the GT-AG rule of Breathnach et al. (1978). The repeated sequences within exons referred to in the text as A and B are indicated by underlining. For each junction, ten intron nucleotides beyond the point at which all eight genes are identical (shaded regions) are shown.

amount of intervening RNA are required for splicing. This suggestion is supported by studies of the sequence requirements for splicing the SV40 late 16S mRNA precursor (Lai and Khoury, 1979; Gruss et al., 1979) which show that deletion of an exon/intron junction or the perfect excision of the entire intron prevents normal mRNA production. In addition, an SV40 deletion mutant in which nine nucleotides 3′ to the leader splice junction are intact produces normally spliced late mRNA, while a deletion mutant in which only seven intron nucleotides remain makes large amounts of unspliced RNA (P. K. Ghosh, J. Mertz, P. Lebowitz and S. Weissman, unpublished results). Finally, Hamer and Leder (1979b) have shown that removal of all but 18 nucleotides of exon 2 of the mouse β^{maj} globin gene still allows normal splicing of IVS2.

The junction sequence alignments, along with the results of computer homology searches, reveal an unusual feature of the globin coding sequences adjacent to the junctions (Figure 8). Within a given gene an 8 bp sequence ("A") at the 3′ end of exon 1 is directly repeated at the 5′ end of exon 3, and another 8 bp sequence ("B") at the 5′ end of exon 2 is also found in exon 3, separated from A by 3 bp. These sequences appear in the mature mRNA as a direct repetition of the unit A-B, in which A and B are separated by 3 bp.

3′ Noncoding and Flanking Regions

Figure 9 shows an alignment of the 3′ noncoding regions [termination codon to the poly(A) addition site] of the human β-like globin genes and the rabbit β and mouse β^{maj} and β^{min} genes. The site of poly(A) addition has been determined only for the human β (Proudfoot, 1977) and γ (Poon, Kan and Boyer, 1978) genes and the rabbit β gene (Proudfoot, 1977). With the exception of the region near the poly(A) addition sites, there is very little homology among the different 3′ noncoding sequences (Table 2). Even the very closely related $^{G}\gamma$ and $^{A}\gamma$ genes differ by 7% in the 3′ noncoding region.

The hexanucleotide AATAAA, first noted by Proudfoot and Brownlee (1976), precedes the poly(A) addition site in all the β-like genes of Figure 9. This sequence is thought to play a role in either or both the processing or polyadenylation of mRNA molecules. For example, T. Shenk et al. (personal communication) have constructed variants of SV40 which carry a tandem duplication of an AATAAA-containing 3′ noncoding segment. Analysis of the species of late RNA produced by these variants indicates that polyadenylation can occur 3′ to either of the duplicated AAUAAA sequences.

As with other eucaryotic genes, the distance between the AATAAA sequence and the poly(A) addition site is not conserved among mammalian β-like globin genes. The sequences immediately adjacent to the poly(A) addition site are relatively conserved among adult β-like genes. These sequences are not present in the human γ- and ε-globin genes, although they do appear in certain other eucaryotic genes (see Figure 8 of Benoist et al., 1980). The sequences on either side of the AATAAA hexanucleotide are homologous in all the β-like genes of Figure 9, except that in the γ-globin genes, part of this region has been deleted. These homologies are not shared with other eucaryotic genes.

In making the alignments of Figure 9, we assumed that there is a large deletion near the center of the 3′ noncoding regions of the human γ genes and a different large deletion in the rabbit β-globin gene. Several smaller deletions (or insertions) are also indicated. The occurrence of sizable deletions or insertions dur-

ing the evolution of these genes suggests that the particular sequences which comprise the 3′ noncoding regions are not essential to globin mRNA function. Consistent with this idea are experimental results showing that the elimination of the 3′ noncoding region of rabbit β-globin mRNA does not abolish in vitro translation (Kronenberg, Roberts and Efstratiadis, 1979), and the observation that a short form of mouse dihydrofolate reductase (DHFR) mRNA lacking the 3′ terminal 850 bases of the 3′ noncoding region is polyadenylated and translatable in vitro (Nunberg et al., 1980). It is noteworthy that the normal DHFR mRNA contains an AAUAAA sequence near its 3′ end, while the short form contains the sequence AUAA in the analogous position.

We were unable to find any strong sequence homologies in the regions located 3′ to the poly(A) addition sites. The sequence TTTT or TTGT occurs six nucleotides 3′ to the poly(A) addition sites of certain eucaryotic genes (see Lai et al., 1979). Similar tetranucleotide sequences are found 2–15 nucleotides into the 3′ flanking regions of all sequenced mammalian β-like globin genes. However, the significance of these homologies is questionable in view of the variability in their sequence and location. Furthermore, we do not find GC-rich palindromes preceding T stretches such as those which occur near the transcription termination sites of procaryotic genes (for review see Rosenberg and Court, 1979). The lack of such sequences in globin and other eucaryotic genes transcribed by RNA polymerase II suggests that processing and polyadenylation may be the mechanism by which the ends of their mRNAs are defined (Fraser et al., 1979). The existence of nuclear transcripts which extend beyond the poly(A) addition sites of adenovirus late mRNAs is consistent with this idea (Nevins and Darnell, 1978; Fraser et al., 1979).

A Mechanism for Generating Deletions within and surrounding β-Like Globin Genes

Pairwise comparisons of noncoding sequences in human, mouse and rabbit β-like globin genes reveal that these regions diverge by deletion and addition as well as by simple base substitution (Konkel et al., 1979;

Figure 9. Alignment of the 3′ Noncoding Sequences of Mammalian β-Like Globin Genes

An alignment of the sequences between the termination codon and the poly(A) addition site of eight mammalian β-like globin genes is shown. Due to the high degree of sequence divergence in this region, the alignment should be considered as representative of a number of equally consistent alignments. The location of the sequence AATAAA, which is common to polyadenylated mRNAs (Proudfoot and Brownlee, 1976), is indicated by shading. Dashes indicate gaps introduced as required for alignment (see text). Similarly, nucleotides shown above a given sequence were displaced to achieve the most consistent alignment of the sequences on either side. Termination codons are underlined and poly(A) addition sites (where known) are indicated by asterisks.

Figure 10 Examples of Deletions Flanked by Short Direct Repeats within Noncoding Sequences of Mammalian β-Like Globin Genes

Several examples of pairwise alignment of sequences within noncoding regions of mammalian β-like globin genes are shown. In these examples, the best alignment of the two sequences is achieved by assuming a deletion in one sequence (upper line of each pair) with respect to the other (lower line). Dashes represent the nucleotides not present in the upper sequence. Short direct repeats near the ends of the deletions are underlined. The two aligned $^A\gamma$ large intron sequences are those of two different alleles (Slightom et al., 1980).

van Ooyen et al., 1979; this paper). Examination of the nucleotide sequences surrounding the putative deletion sites strongly suggests that short (2–8 bp) direct repeats may be involved in the generation of these deletions. Several examples of this phenomenon are shown in Figure 10. The best alignment of the noncoding sequences of a number of different gene pairs is achieved by assuming that a deletion occurred in one of the genes. This assumption is justified by the observation that in alignments such as those shown in Figures 4, 6 and 9, genes which are missing a particular part of the aligned sequence are in the minority. In Figure 10, we have placed the sequence containing the putative deletion on the top line of each pairwise comparison. In every case we can locate a direct repeat near the endpoints of the deletion. The deletion removes one of the repeats entirely and either none or part of the other repeat. This pattern is remarkably similar to that observed by Farabaugh and Miller (1978) for deletions in the *lac I* gene of E. coli in which hotspots for spontaneous mutations are flanked by short direct repeat sequences. Farabaugh and Miller (1978) point out that the presence of direct repeats could promote deletions by slipped mispairing during DNA replication according to a model proposed by Streisinger et al. (1966) for the generation of frameshift mutations.

The deletions observed in β-like globin genes are unlikely to be a consequence of unequal crossing over between repeat sequences, for two reasons. First, most of the repeat sequences are very short. Second, an unequal crossing over event would produce one daughter molecule with three intact copies of the repeat sequence and another molecule with one intact copy. Only one example of the former and very few examples of the latter were observed in our analysis.

Figure 11 shows a model for the generation of deletions within the β-like globin genes. Figure 11A shows a region of duplex DNA containing direct repeat sequences, which are labeled R1 and R2 on one strand and R1′ and R2′ on the other. As the replication fork moves through this region, the duplex containing the repeat sequences will become single-stranded (Figure 11B). More than 1 kb of single-stranded DNA can be detected in the replication fork of bacteriophage T7 DNA (Wolfson and Dressler, 1972, 1979). Once the two strands are separated, the R2 repeat would be free to base-pair with the complementary R1′ sequence (Figure 11C). Such an event would produce a single-stranded loop containing the R1 repeat as well as the sequences which lie between R1 and R2. This loop could then be recognized by DNA repair enzymes which would excise the loop and rejoin the ends of the broken DNA strand (Figure 11D). Subsequent replication would generate a normal daughter duplex and a duplex containing only one of the two repeats and lacking the sequences between

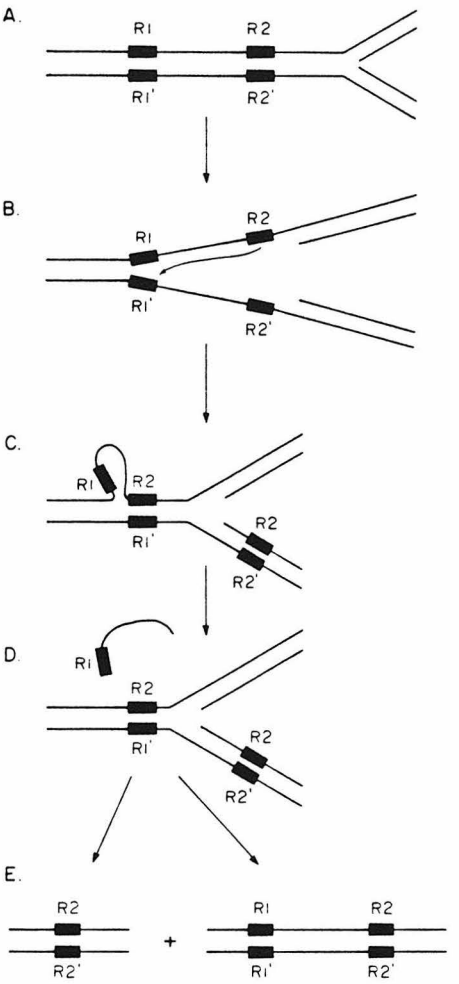


Figure 11. A "Slipped Mispairing" Model for the Generation of Deletions during DNA Replication

The diagram shows successive stages in the passage of a replication fork through a region of DNA containing two short direct repeats. Solid lines represent strands of DNA. The short direct repeats (R1 and R2), as well as their complements (R1′ and R2′), are represented by solid boxes. See text for description of the model.

R1 and R2 (Figure 11E). If the excised DNA loop includes part of the R2 repeat, the daughter duplex will contain only a truncated R2 sequence. On the other hand, if the excised loop includes R1 and only a portion of the region between R1 and R2, the daughter duplex will contain R2 and part of the sequence between the repeats. Thus all the putative

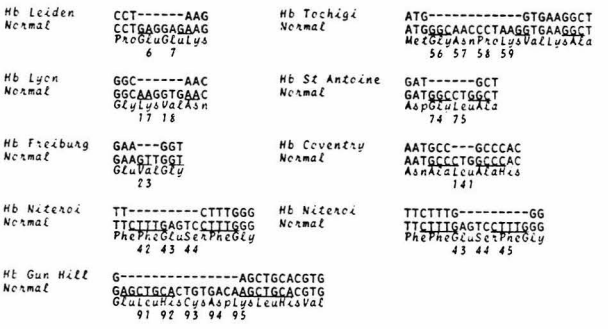


Figure 12. Short Direct Repeats Flanking Known Deletions in the Coding Sequences of the Human β-Globin Gene

Several known deletion variants of human β-globin (Bunn et al., 1977), in which one or more amino acids are missing from the polypeptide chain, are interpreted in terms of their presumed gene sequence. In each comparison, the sequence of the appropriate portion of the normal human β-globin gene is shown with the normal amino acid sequence under it. Numbers refer to the amino acids deleted in the variant. The presumed gene sequence of the variant is shown on the upper line of each comparison, with dashes representing the deleted nucleotides. Short direct repeats occurring near the ends of each deletion are underlined in the normal sequence.

deletion events illustrated in Figure 10 can be accounted for by variability in the excision of the single-stranded loop.

Well characterized deletions exist in the coding sequence of the human β-globin gene. These deletions were identified by amino acid sequence analysis of structural hemoglobin variants (Bunn et al., 1977). Marotta et al. (1977) noted the presence of short, direct repeat sequences near the endpoints of these deletions and suggested alternative mechanisms for the participation of repeats in the deletion event. The coding sequence deletions can also be accounted for by the model described above. In Figure 12, we have aligned the normal human β-globin sequence (bottom line) with the DNA sequence predicted by the amino acid sequence of the structural variants. Based on these alignments and those of Figure 10, it seems probable that many deletion mutations in eucaryotic genes are promoted by the presence of short, direct repeat sequences.

Acknowledgments

We are grateful to other members of our laboratories for valuable discussion and useful suggestions, and to Connie Katz for her tireless efforts in preparing the manuscript. We thank R. J. Britten for assistance in the dot matrix analysis, J. Bonner for the use of his computer, and J. Lingrel for making available unpublished data.

Work in our laboratories was supported by grants from the NIH, the NSF and the British Medical Research Council. In addition, A. E., J. W. P. and T. M. were supported, respectively, by a Basil O'Connor starter grant from the March of Dimes Birth Defects Foundation, by an NIH predoctoral training grant to the California Institute of Technology and by the Rita Allen Foundation. Part of the cost of preparing the manuscript was defrayed by the W. J. Sloan Foundation of the California Institute of Technology.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received June 18, 1980

References

Baker, C. C., Herisse, J., Courtois, G., Galibert, F. and Ziff, E. (1979). Messenger RNA for the Ad2 DNA binding protein: DNA sequences encoding the first leader and heterogeneity at the mRNA 5′ end. Cell *18*, 569–580.

Bakker, R. T. (1975). Dinosaur renaissance. Sci. Am. *232*, 57–78.

Baralle, F. E. and Brownlee, G. G. (1978). AUG is the only recognisable signal sequence in the 5′ non-coding regions of eukaryotic mRNA. Nature *274*, 84–87.

Baralle, F. E., Shoulders, C. C. and Proudfoot, N. J. (1980). The primary structure of the human ε-globin gene. Cell *21*, 621–626.

Bell, G. I., Pictet, R. L., Rutter, W. J., Cordell, B., Tischer, E. and Goodman, H. M. (1980). Sequence of the human insulin gene. Nature *284*, 26–32.

Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980). The ovalbumin gene—sequence of putative control regions. Nucl. Acids Res. *8*, 127–142.

Bernards, R., Little, P. F. R., Annison, G., Williamson, R. and Flavell, R. A. (1979). Structure of the human $^{G}\gamma$-$^{A}\gamma$-δ-β-globin gene locus. Proc. Nat. Acad. Sci. USA *76*, 4827–4831.

Blake, C. C. F. (1979). Exons encode protein functional units. Nature *277*, 598.

Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P. (1978). Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. Proc. Nat. Acad. Sci. USA *75*, 4853–4857.

Bunn, H. F., Forget, B. G. and Ranney, H. M. (1977). Human Hemoglobins (Philadelphia: W. B. Saunders).

Chang, J. C., Poon, R., Neumann, K. H. and Kan, Y. W. (1978). The nucleotide sequence of the 5′ untranslated region of human γ-globin mRNA. Nucl. Acids Res. *5*, 3515–3522.

Cordell, B., Bell, G., Tischer, E., DeNoto, F. M., Ullrich, A., Pictet, R., Rutter, W. J. and Goodman, H. M. (1979). Isolation and characterization of a cloned rat insulin gene. Cell *18*, 533–543.

Craik, C. S., Buchman, S. R. and Beychok, S. (1980). Characterization of globin domains: heme binding to the central exon product. Proc. Nat. Acad. Sci. USA *77*, 1384–1388.

Dayhoff, M. O. (1972). Atlas of Protein Sequence and Structure (Washington, D.C.: National Biomedical Research Foundation).

Deisseroth, A., Nienhuis, A., Turner, P., Velez, R., Anderson, W. F., Lawrence, J., Creagan, R. and Kucherlapati, R. (1977). Localization of the human α-globin structural gene to chromosome 16 in somatic cell hybrids by molecular hybridization assay. Cell *12*, 205–218.

Deisseroth, A., Nienhuis, A., Lawrence, J., Giles, R., Turner, P. and Ruddle, F. H. (1978). Chromosomal localization of human β-globin gene on human chromosome 11 in somatic cell hybrids. Proc. Nat. Acad. Sci. USA *75*, 1456–1460.

Dickerson, R. E. (1971). The structure of cytochrome c and the rates of molecular evolution. J. Mol. Evol. *1*, 26–45.

Dickerson, R. E. and Geis, I. (1980). Proteins: Structure, Function and Evolution (Menlo Park, California: Benjamin/Cummings Publishing), in press.

Eaton, W. A. (1980). The relationship between coding sequences and function in hemoglobin. Nature *284*, 183–185.

Embury, S., Lebo, R., Dozy, A. and Kan, Y. W. (1979). Organization of the α-globin genes in the Chinese α-thalassemia syndromes. J. Clin. Invest. *63*, 1307–1310.

Farabaugh, P. J. and Miller, J. H. (1978). Genetic studies of the lac repressor. VII. On the molecular nature of spontaneous hotspots in the lac I gene of Escherichia coli. J. Mol. Biol. *126*, 847–863.

Flavell, R. A., Kooter, J. M., De Boer, E., Little, P. F. R. and Williamson, R. (1978). Analysis of the β–δ–globin gene loci in normal and Hb Lepore DNA: direct determination of gene linkage and intergene distance. Cell *15*, 25–41.

Forget, B. G., Cavallesco, C., DeRiel, J. K., Spritz, R. A., Choudary, P. V., Wilson, J. T., Wilson, L. B., Reddy, V. B. and Weissman, S. M. (1979). Structure of the human globin genes. In Eucaryotic Gene Regulation, ICN-UCLA Symposium on Molecular and Cellular Biology, *XIV*, R. Axel, R. Maniatis and C. F. Fox, eds. (New York: Academic Press), pp. 367–381.

Fraser, N. W., Nevins, J. R., Ziff, E. and Darnell, J. E., Jr. (1979). The major late adenovirus type-2 transcription unit: termination is downstream from the last poly(A) site. J. Mol. Biol. *129*, 643–656.

Fritsch, E. F., Lawn, R. M. and Maniatis, T. (1979). Characterisation of deletions which affect the expression of fetal globin genes in man. Nature *279*, 598–603.

Fritsch, E. F., Lawn, R. M. and Maniatis, T. (1980). Molecular cloning and characterization of the human β-like globin gene cluster. Cell *19*, 959–972.

Gilbert, W. (1979). Introns and exons: playgrounds of evolution. In Eukaryotic Gene Regulation, ICN-UCLA Symposium on Molecular and Cellular Biology, *XIV*, R. Axel, T. Maniatis and C. F. Fox, eds. (New York: Academic Press), pp. 1–12.

Goldberg, M. (1979). Ph.D. thesis, Stanford University, Stanford, California.

Grosschedl, R. and Birnstiel, M. L. (1980). Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants in vivo. Proc. Nat. Acad. Sci. USA *77*, 1432–1436.

Gruss, P., Lai, C. J., Dhar, R. and Khoury, G. (1979). Splicing as a requirement for biogenesis of functional 16S mRNA of simian virus 40. Proc. Nat. Acad. Sci. USA *76*, 4317–4321.

Gusella, J., Varsanyi-Breiner, A., Kao, F.-T., Jones, C., Puck, T. T., Keys, C., Orkin, S. and Housman, D. (1979). Precise localization of human β-globin gene complex on chromosome 11. Proc. Nat. Acad. Sci. USA *76*, 5239–5243.

Hamer, D. H. and Leder, P. (1979a). Expression of the chromosomal mouse β^{maj}-globin gene cloned in SV40. Nature *281*, 35–40.

Hamer, D. H. and Leder, P. (1979b). SV40 recombinants carrying a functional RNA splice junction and polyadenylation site from the chromosomal mouse β^{maj} globin gene. Cell *17*, 737–747.

Hardison, R. C., Butler, E. T., III, Lacy, E., Maniatis, T., Rosenthal, N. and Efstratiadis, A. (1979). The structure and transcription of four linked rabbit β-like globin genes. Cell *18*, 1285–1297.

Haynes, J. R., Rosteck, P. and Lingrel, J. B. (1980). Unusual sequence homology at the 5′ ends of the developmentally regulated β^A, β^C and γ globin genes of the goat. Proc. Nat. Acad. Sci. USA, in press.

Heindell, H. C., Liu, A., Paddock, G. V., Studnicka, G. M. and Salser, W. A. (1978). The primary sequence of rabbit α-globin mRNA. Cell *15*, 43–54.

Hentschel, C., Irminger, J.-C., Bucher, P. and Birnstiel, M. L. (1980). Sea urchin histone mRNA termini are located in gene regions downstream from putative regulatory sequences. Nature *285*, 147–151.

Holmquist, R. (1972). Theoretical foundations for a quantitative approach to paleogenetics. J. Mol. Evol. *1*, 115–133.

Jahn, C. L., Hutchison, C. A., III, Phillips, S. J., Weaver, S., Haigwood, N. L., Voliva, C. F. and Edgell, M. H. (1980). DNA sequence organization of the β-globin complex in the BALB/c mouse. Cell *21*, 159–168.

Jeffreys, A. J. (1979). DNA sequence variants in the $^{G}\gamma$–, $^{A}\gamma$–, δ– and β–globin genes of man. Cell *18*, 1–10.

Jeffreys, A. J. and Flavell, R. A. (1977). A physical map of the DNA regions flanking the rabbit β–globin gene. Cell *12*, 429–439.

Jeffreys, A., Craig, I. and Francke, U. (1979). Localisation of the $^{G}\gamma$–, $^{A}\gamma$–, δ– and β–globin genes on the short arm of human chromosome 11. Nature *281*, 606–608.

Kantor, J. A., Turner, P. H. and Nienhuis, A. W. (1980). Beta thalassemia: mutations which affect processing of the β–globin mRNA precursor. Cell *21*, 149–157.

Kimura, M. and Ohta, T. (1972). On the stochastic model for estimation of mutational distance between homologous proteins. J. Mol. Evol. *2*, 87–90.

Konkel, D. A., Tilghman, S. M. and Leder, P. (1978). The sequence of the chromosomal mouse β–globin major gene. homologies in capping, splicing and poly(A) sites. Cell *15*, 1125–1132.

Konkel, D. A., Maizel, J. V., Jr. and Leder, P. (1979). The evolution and sequence comparison of two recently diverged mouse chromosomal β–globin genes. Cell *18*, 865–873.

Kozak, M. (1978). How do eucaryotic ribosomes select initiation regions in messenger of RNA? Cell *15*, 1109–1123.

Kronenberg, H. M., Roberts, B. E. and Efstratiadis, A. (1979). The 3′ noncoding region of the β–globin mRNA is not essential for in vitro translation. Nucl. Acids Res. *6*, 153–166.

Lacy, E., Hardison, R. C., Quon, D. and Maniatis, T. (1979). The linkage arrangement of four rabbit β–like globin genes. Cell *18*, 1273–1283.

Lai, C.-J. and Khoury, G. (1979). Deletion mutants of simian virus 40 defective in biosynthesis of late viral mRNA. Proc. Nat. Acad. Sci. USA *76*, 71–75.

Lai, E. C., Stein, J. P., Catterall, J. F., Woo, S. L. C., Mace, M. L., Means, A. R. and O'Malley, B. W. (1979). Molecular structure and flanking nucleotide sequences of the natural chicken ovomucoid gene. Cell *18*, 829–842.

Lauer, J., Shen, C.-K. J. and Maniatis, T. (1980). The chromosomal arrangement of human α–like globin genes: sequence homology and α–globin gene deletions. Cell *20*, 119–130.

Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G. and Maniatis, T. (1978). The isolation and characterization of linked δ– and β–globin genes from a cloned library of human DNA. Cell *15*, 1157–1174.

Lawn, R. M., Efstratiadis, A., O'Connell, C. and Maniatis, T. (1980). The nucleotide sequence of the human β–globin gene. Cell *21*, 647–651.

Lebo, R., Carrano, A., Burkhart-Schultz, K., Dozy, A., Yu, L. C. and Kan, Y. W. (1979). Assignment of human β–, γ–, and δ–globin genes to the short arm of chromosome 11 by chromosome sorting and DNA restriction enzyme analysis. Proc. Nat. Acad. Sci. USA *76*, 5804–5808.

Lerner, M., Boyle, J., Mount, S., Wolin, S. and Steitz, J. (1980). Are snRNPs involved in splicing? Nature *283*, 220–224.

Liebhaber, S. A., Goosens, M. and Kan, Y. W. (1980). Cloning and complete nucleotide sequence of the human 5′ α globin gene. Proc. Nat. Acad. Sci. USA, in press.

Little, P. F. R., Flavell, R. A., Kooter, J. M., Annison, G. and Williamson, R. (1979a). Structure of the human fetal globin gene locus. Nature *278*, 227–231.

Little, P. F. R., Williamson, R., Annison, G., Flavell, R. A., De Boer, E., Bernini, L. G., Ottolenghi, S., Saglio, G. and Mazza, U. (1979b). Polymorphisms of human γ–globin genes in Mediterranean populations. Nature *282*, 316–318.

McKenna, M. C. (1969). The origin and early differentiation of therian mammals. Ann. NY Acad. Sci. *167*, 217–240.

Maniatis, T., Fritsch, E. F., Lauer, J. and Lawn, R. M. (1980). The

molecular genetics of human hemoglobins. Ann. Rev. Genet., in press.

Manley, J. L., Fire, A., Cano, A., Sharp, P. A. and Gefter, M. L. (1980). DNA-dependent transcription of adenovirus genes in a soluble whole-cell extract. Proc. Nat. Acad. Sci. USA 77, 3855–3859.

Mantei, N., Boll, W. and Weissman, C. (1979). Rabbit β-globin mRNA production in mouse L cells transformed with cloned rabbit β-globin chromosomal DNA. Nature 281, 40–46.

Marotta, C. A., Wilson, J. T., Forget, B. G. and Weissman, S. M. (1977). Human β-globin messenger RNA. III. Nucleotide sequences derived from complementary DNA. J. Biol. Chem. 252, 5040–5053.

Mears, J. G., Ramirez, F., Leibowitz, D. and Bank, A. (1978). Organization of human δ- and β-globin genes in cellular DNA and the presence of intragenic inserts. Cell 15, 15–23.

Moore, G. W., Goodman, M., Callahan, C., Holmquist, R. and Moise, H. (1976). Stochastic versus augmented maximum parsimony method for estimating superimposed mutations in the divergent evolution of protein sequences. Methods tested on cytochrome c amino acid sequences. J. Mol. Biol. 105, 15–37.

Nevins, J. R. and Darnell, J. E., Jr. (1978). Steps in the processing of Ad2 mRNA: poly(A)$^+$ nuclear sequences are conserved and poly(A) addition precedes splicing. Cell 15, 1477–1493.

Nishioka, Y. and Leder, P. (1979). The complete sequence of a chromosomal mouse α-globin gene reveals elements conserved throughout vertebrate evolution. Cell 18, 875–882.

Nunberg, J. H., Kaufman, R. J., Chang, A. C. Y., Cohen, S. N. and Schimke, R. T. (1980). Structure and genomic organization of the mouse dihydrofolate reductase gene. Cell 19, 355–364.

Orkin, S. H. (1978). The duplicated human alpha globin genes lie close together in cellular DNA. Proc. Nat. Acad. Sci. USA 75, 5950–5954.

Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. and Dodgson, J. (1980). The evolution of genes: the chicken preproinsulin gene. Cell 20, 555–565.

Poon, R., Kan, Y. W. and Boyer, H. W. (1978). Sequence of the 3′ noncoding and adjacent coding regions of human γ-globin mRNA. Nucl. Acids Res. 5, 4625–4630.

Pribnow, D. (1979). Genetic control signals in DNA. In Biological Regulation and Development, 1. R. Goldberger, ed. (New York: Plenum Press), pp. 219–277.

Proudfoot, N. J. (1977). Complete 3′ noncoding region sequences of rabbit and human β-globin messenger RNAs. Cell 10, 559–570.

Proudfoot, N. J. and Brownlee, G. G. (1976). 3′ non-coding region sequences in eukaryotic messenger RNA. Nature 263, 211–214.

Proudfoot, N. and Baralle, F. (1979). Molecular cloning of the human ε-globin gene. Proc. Nat. Acad. Sci. USA 76, 5435–5439.

Proudfoot, N. J. and Maniatis, T. (1980). The structure of a human α-globin pseudogene and its relationship to α-globin gene duplication. Cell 21, 537–544.

Ramirez, F., Burns, A. L., Mears, J. G., Spence, S., Starkman, D. and Bank, A. (1979). Isolation and characterization of cloned human fetal globin genes. Nucl. Acids Res. 7, 1147–1162.

Richards, R. I., Shine, J., Ullrich, A., Wells, J. R. E. and Goodman, H. W. (1979). Molecular cloning and sequence analysis of adult chicken β-globin cDNA. Nucl. Acids Res. 7, 1137–1146.

Romero-Herrera, A. E., Lehmann, H., Joysey, K. A. and Friday, A. E. (1973). Molecular evolution of myoglobin and the fossil record: a phylogenetic synthesis. Nature 246, 389–395.

Rosenberg, M. and Court, D. (1979). Regulatory sequences involved in the promotion and termination of RNA transcription. Ann. Rev. Genet. 13, 319–353.

Salser, W. (1978). Globin mRNA sequences: analysis of base pairing and evolutionary implications. Cold Spring Harbor Symp. Quant. Biol. 42, 985–1002.

Salser, W. A., Cummings, I., Liu, A., Strommer, J., Padayatty, J. and Clarke, P. (1979). Analysis of chicken globin cDNA clones: discovery of a novel chicken alpha-globin gene induced by stress in young chickens. In Cellular and Molecular Regulation of Hemoglobin Switching, G. Stamatoyannopoulos and A. W. Nienhuis, eds. (New York: Grune and Stratton), pp. 621–645.

Schaffner, W., Kunz, G., Daetwyler, H., Telford, J., Smith, H. O. and Birnstiel, M. L. (1978). Genes and spacers of cloned sea urchin histone DNA analyzed by sequencing. Cell 14, 655–671.

Seif, I., Khoury, G. and Dhar, R. (1979). BKV splice sequences based on analysis of preferred donor and acceptor sites. Nucl. Acids Res. 6, 3387–3398.

Slightom, J. L., Blechl, A. E. and Smithies, O. (1980). Human fetal ^{G}γ- and ^{A}γ-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell 21, 627–638.

Smithies, O., Blechl, A., Denniston-Thompson, K., Newell, N., Richards, J. E., Slightom, J. L., Tucker, P. W., and Blattner, F. R. (1978). Cloning human fetal γ-globin and mouse α-type globin DNA: characterization and partial sequencing. Science 202, 1284–1289.

Spritz, R. A., DeRiel, J. K., Forget, B. G. and Weissman, S. M. (1980). Complete nucleotide sequence of the human δ-globin gene. Cell 21, 639–646.

Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E. and Inouye, M. (1966). Frameshift mutations and the genetic code. Cold Spring Harbor Symp. Quant. Biol. 31, 77–84.

Sures, I., Lowry, J. and Kedes, L. H. (1978). The DNA sequence of sea urchin (S. purpuratus) H2A, H2B and H3 histone coding and spacer regions. Cell 15, 1033–1044.

Sures, I., Levy, S. and Kedes, L. H. (1980). Leader sequences of Strongylocentrotus purpuratus histone mRNAs start at a unique heptanucleotide common to all five histone genes. Proc. Nat. Acad. Sci. USA 77, 1265–1269.

Tilghman, S. M., Tiemeier, D. C., Seidman, J. G., Peterlin, B. M., Sullivan, M., Maizel, J. and Leder, P. (1978). Intervening sequence of DNA identified in the structural portion of a mouse β-globin gene. Proc. Nat. Acad. Sci. USA 75, 725–729.

Tsujimoto, Y. and Suzuki, Y. (1979). Structural analysis of the fibroin gene at the 5′ end and its surrounding regions. Cell 16, 425–436.

Tuan, D., Biro, P. A., DeRiel, J. K., Lazarus, H. and Forget, B. G. (1979). Restriction endonuclease mapping of the human γ-globin gene loci. Nucl. Acids Res. 6, 2519–2544.

van Ooyen, A., van den Berg, J., Mantei, N. and Weissman, C. (1979). Comparison of total sequence of a cloned rabbit β-globin gene and its flanking regions with a homologous mouse sequence. Science 206, 337–344.

Van Ormondt, H., Maat, J., de Waard, A. and Van der Eb, A. J. (1978). The nucleotide sequence of the transforming HpaI-E fragment of adenovirus type 5 DNA. Gene 4, 309–328.

Wasylyk, B., Kédinger, C., Corden, J., Brison, O. and Chambon, P. (1980). Specific in vitro initiation of transcription on conalbumin and ovalbumin genes and comparison with adenovirus-2 early and late genes. Nature 285, 367–373.

Weatherall, D. J. and Clegg, J. B. (1979). Recent developments in the molecular genetics of human hemoglobin. Cell 16, 467–479.

Weil, P. A., Luse, D. S., Segall, J. and Roeder, R. G. (1979). Selective and accurate initiation of transcription at the Ad2 major late promoter in a soluble system dependent on purified RNA polymerase II and DNA. Cell 18, 469–484.

Wigler, M., Sweet, R., Sim, G. K., Wold, B., Pellicer, A., Lacy, E., Maniatis, T., Silverstein, S. and Axel, R. (1979). Transformation of mammalian cells with genes from procaryotes and eucaryotes. Cell 16, 777–785.

Wilson, A. C., Carlson, S. S. and White, T. J. (1977). Biochemical evolution. Ann. Rev. Biochem. 46, 573–639.

Wolfson, J. and Dressler, D. (1972). Regions of single-stranded DNA in the growing points of replicating bacteriophage T7 chromosomes. Proc. Nat. Acad. Sci. USA 69, 2682–2686.

Wolfson, J. and Dressler, D. (1979). Bacteriophage T7 DNA replication. J. Biol. Chem. 254, 10490–10495.