J Mol Evol (1984) 20:195–201

Journal of
Molecular Evolution

Insertion of a Short Repetitive Sequence (D88I) in a Sea Urchin Gene: A Typical Interspersed Repeat?

Steven A. Johnson[1], Eric H. Davidson[1], and Roy J. Britten[1,2]

[1] Division of Biology, California Institute of Technology, Pasadena, California 91125, USA
[2] Carnegie Institution of Washington, Washington, DC 20015, USA

Summary. A comparison has been made between the Sp88 gene regions of the DNAs of the sea urchins *Strongylocentrotus purpuratus* (Sp.) and *Strongylocentrotus drobachiensis* (Sd.). Examination of the 3′ terminal part of the transcribed region revealed a short repetitive sequence present in Sd. but absent from Sp. A 12-nucleotide sequence present once in Sp. is almost perfectly duplicated at both ends of the repeat in Sd., suggesting that a mobile repeat was inserted in the Sd. genome. Other members of this family of repeated sequences occur in many interspersed locations in the genomes of both species. Except for the insertion duplication, the inserted sequence lacks direct or reverse repeats.

Key words: DNA—Mobile genetic element—Enhancer—Rare maternal transcript

Introduction

It has been known for a number of years that the genomic frequencies of interspersed repeated DNA sequences have changed during sea urchin evolution (Moore et al. 1978, 1981). Significant evolutionary frequency changes have been observed for individual families of repeats and for the whole set of sea urchin repeats. Since short repeats are interspersed throughout the genomes of these species, it is clear that frequent insertion events involving short repeats have occurred during their evolution. It can be argued that this is a generally occurring process in the evolution of eukaryotic genomes (Britten 1982).

Offprint requests to: R.J. Britten

Much more is known about *Drosophila* movable elements than about such elements in other animal DNAs. Their insertion is a major source of mutation (Rubin 1983). We do not yet know if an analogous situation occurs in other eukaryotes. Striking features distinguish the *Drosophila* genome from that of typical eukaryotes. It is more than an order of magnitude smaller and has a long-period pattern of interspersion of repeats and single-copy DNA, compared with the typical short-period pattern of most eukaryotes (Davidson and Britten 1979). In addition, we know that massive amounts of reorganization have occurred in dipteran evolution, since the housefly has a more normal genome size and a short-period interspersion pattern. In comparison, mammals representing a comparable period of evolution all have very similar genome sizes and probably all show the short-period pattern. Thus, it would not be surprising to find significant differences between mammals, dipterans, and sea urchins in the types of mobile genetic elements and their rates of insertion and deletion.

We are currently attempting to determine the connection between repeated sequences and mobile repetitive elements. We report here the identification, by interspecies DNA comparison, of a single repeated sequence that has been inserted into the sea urchin Sp88 gene region, which has previously been studied in some detail (Thomas et al. 1982). This gene is represented by a typical low-prevalence maternal poly(A) RNA that is 9.5 kb long and contains several repetitive sequences. The putative mobile repetitive sequence being reported here was inserted in the *Strongylocentrotus drobachiensis* (Sd.) genome in an apparently noncoding DNA region about 1.6 kb upstream of the 3′ terminus of the transcript of the Sp88 gene.

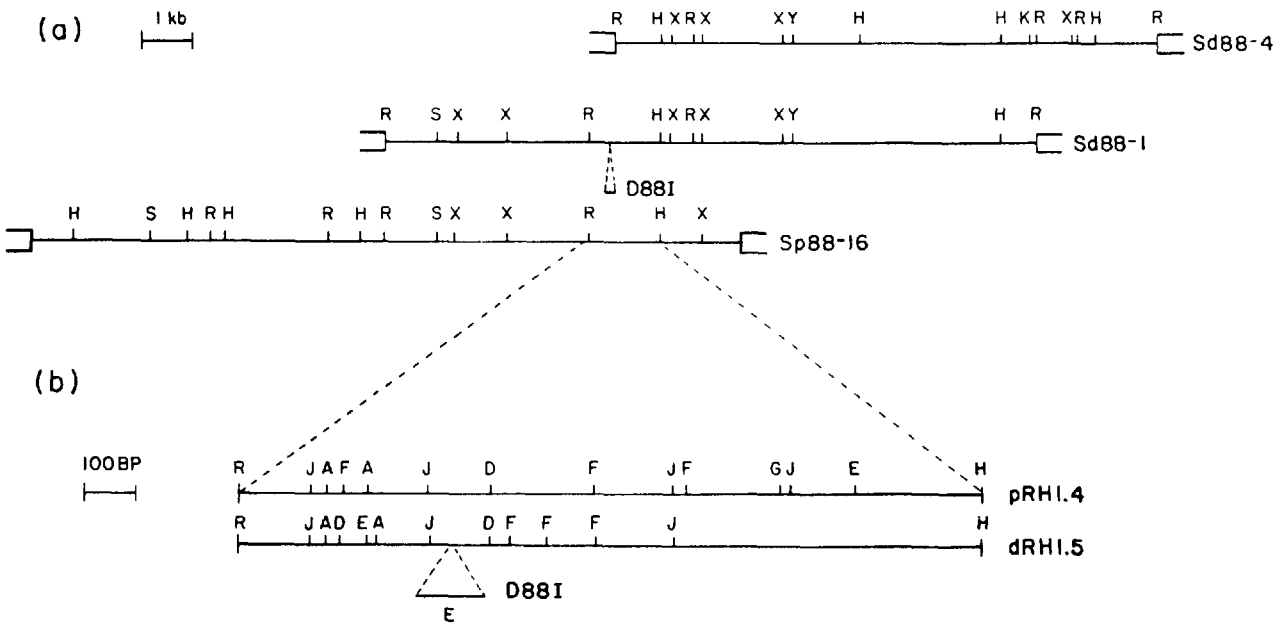

Fig. 1a, b. **a** Comparison of restriction maps of Sd88-1, Sd88-4, and Sp88-16 recombinant lambda DNAs. The termini of Sp88-16 are artificial EcoR1 sites attached for cloning and thus would not be expected in Sd88-1 and Sd88-4. The D88I insertion has been removed from Sd88-1 to allow comparison with sites within Sd88-4 and Sp88-16. The map for Sp88-16 is taken from Thomas et al. (1982). Restriction enzymes used were as follows: R, EcoR1; H, HindIII; X, XbaI; Y, XhoI; S, SalI; K, KpnI. **b** Four-base-site restriction map of the 1.4-kb EcoR1–HindIII fragment pRH1.4 and its Sd. homologue, dRH1.5. The D88I insertion element has been removed to facilitate comparison of the two maps. Restriction enzymes used were as follows: A, AluI; D, DdeI; E, HaeIII; F, HinfI; G, HpaII; J, RsaI

Materials and Methods

Detailed procedures for Charon 4A library construction, library screening, agarose gel electrophoresis, Southern blotting, and various ancillary methods have been presented elsewhere (Anderson et al. 1981; Thomas et al. 1982). DNA sequencing by the method of Maxam and Gilbert (1977) was done as described by Posakony et al. (1981).

S. drobachiensis DNA Library. For the Charon 4A library construction, long, native DNA from an individual sperm preparation was partially cut with EcoR1. The restricted DNA from a series of partial digestions in which both incubation time and enzyme amount were varied was combined and centrifuged through sucrose gradients. Fractions containing 15–20 kb long DNA were pooled and ligated to purified EcoR1-cut Charon 4A arms and packaged. The total library contained in excess of 3×10^7 plaque-forming units, and was not amplified before screening.

Hybridization Criterion. When gel blots of restricted Sd. recombinant lambda DNA were probed with labeled *Strongylocentrotus purpuratus* (Sp.) fragments, the incubation temperature was either 64° or 60°C to accommodate the average 7% divergence between Sd. and Sp. single-copy DNA (Hall et al. 1980).

Results

Isolation of the S. drobachiensis 88 Gene Region and Interspecies Comparison

Much of the Sp88 gene region contains closely spaced repeats. Nevertheless, a moderately long stretch of uninterrupted single-copy DNA is located near the 3′ end of the proposed transcript (Thomas et al. 1982). A 1.4-kb EcoR1–HindIII fragment from this region has been subcloned (Thomas et al. 1982). This cloned Sp. DNA fragment, pRH1.4, was used to screen a Charon 4A recombinant library made from sperm DNA of a single Sd. individual. Seven positives were obtained from approximately ten genome equivalents of plaques. DNA was prepared from four of these.

Restriction mapping showed that these recombinants represented two allelic length variants for the EcoR1–HindIII fragment. The existence of these two variant lengths in the DNA of this individual had already been recognized from genome blots using the pRH1.4 probe (see Fig. 2a). The restriction maps of two recombinants representing the two alleles are shown in Fig. 1a.

The region of the Sp88 gene represented in Sd88-1 does not appear to have diverged extensively from that of the last common ancestor of Sp. and Sd. Seven of nine restriction sites in the 7-kb region of overlap are common to both species. The Sd. library was formed from a partial EcoR1 digest and the Sp88 clone is from a partial HaeIII, EcoR1 linker library. Thus, the natural EcoR1 site at the left end of the overlap is conserved, whereas the right end is terminated by an artificial site. The median single-copy sequence divergence between Sd. and Sp. is

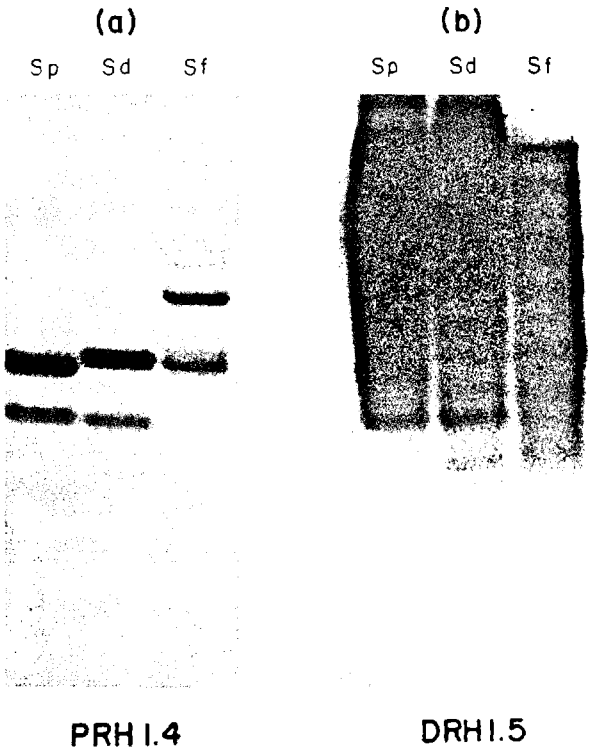

Fig. 2a, b. D88I is an interspersed repeat in Sd., Sp., and Sf. genomic DNA. Total genomic DNA isolated from sperm of an Sp., Sd., or Sf. individual was restricted with EcoR1 and HindIII. Half of the restricted DNA from each reaction was loaded into adjacent lanes on the left side of a 1% neutral agarose gel, and then the loading pattern was repeated on the right side of the same gel using the other half. After electrophoresis and transfer to nitrocellulose, the blot was cut in half. One half was hybridized with nick-translated pRH1.4 (**a**); the other half was hybridized with nick-translated dRH1.5 (**b**). The hybridization criterion for each half was 4×SET (1× SET = 0.15 M NaCl, 0.03 M Tris, pH 8.0, 0.002 M EDTA) at 60°C. The final wash condition for each half was 0.3× SET at 60°C

about 7% (Hall et al. 1980). The expectation is thus that $100 \times (1 - 0.93^6) = 35\%$ of sites would have been lost and an equal number of new sites gained. Therefore, we would expect that in a typical region the majority of restriction sites would differ.

When the two allelic variant Sd. recombinants (Sd88-1 and Sd88-4) are compared, three restriction site differences and eight common sites are seen in the region of overlap. This result is not surprising since single-copy DNA sequence variation averages 2% between the DNAs of Sd. individuals (Grula et al. 1982); thus we expect that about 80% of six-base restriction sites would be held in common between individuals.

A major purpose of this work was to search for rearrangement events such as the insertion or deletion of repeated sequences or mobile elements. As a first step, labeled Sp. genomic DNA was hybridized to Southern blots of several restriction digests of the three recombinants Sp88-16, Sd88-1, and Sd88-4. Repeats of more than 100 copies can be detected in this way, with intensity differences reflecting frequency differences. Homologous fragments were identified from the maps in Fig. 1a; the results (data not shown) indicated no differences in the intensity of the observed hybridization between them. Thus, no rearrangement events were detected by this method, a result supported by the map alignments and fragment sizes, with one interesting exception described in the next section.

Restriction Fragment Size Differences Reveal an Insertion Element

To initiate a more detailed interspecies comparison, we chose the previously cloned 1.4-kb region represented by pRH1.4. The corresponding region from Sd. was cloned in pBR322 from an EcoR1 digest of the Sd88-1 lambda recombinant insert. A 2.1-kb EcoR1 fragment was recovered that contained the Sd. homolog (dRH1.5) (see Fig. 1a).

The maps of pRH1.4 and its Sd. homolog dRH1.5(Fig. 1b) show that a small piece of DNA present in dRH1.5 is absent in pRH1.4. If 131 nucleotides are omitted from the dRH1.5 map, then four additional restriction sites come into alignment, including the right terminal HindIII site, yielding the same length for the remainder of dRH1.5 as for pRH1.4. Therefore, it appears that a short sequence element present in the Sd. genome is absent from Sp. We have termed this fragment D88I to identify it as an insertion in the Sd. 88 gene region.

In characterizing D88I we first asked whether it was repeated. A genome blot was prepared with duplicate lanes containing individual Sp., Sd., and *S. franciscanus* (Sf.) DNA cut with EcoR1 and HindIII. One-half of the blot was hybridized with nick-translated pRH1.4 and the other, identical half was hybridized with nick-translated dRH1.5. The result, presented in Fig. 2, shows that dRH1.5 contains a fragment of DNA that is repeated in Sp., Sd., and Sf. genomic DNA. It is also apparent that pRH1.4 is made up of sequences that are single copy in the Sp., Sd., and Sf. genomes.

A more rigorous proof that the extra element in dRH1.5 was the repeat was obtained as follows. Gel-purified dRH1.5 was cut with Dde1 and Rsal and then fill-in labeled with the Klenow fragment of DNA polymerase 1. Each end-labeled Dde/Rsa fragment was then gel purified and hybridized to individual Southern blot lanes containing HindIII-restricted single-individual genomic Sp. DNA. All but one of the fragments hybridized to a single band. Only the 265-nucleotide (NT) Dde/Rsa fragment (corresponding to the proposed location of the insert) hybridized to a smear of fragments, indicating the presence of a medium- or high-frequency repeat.

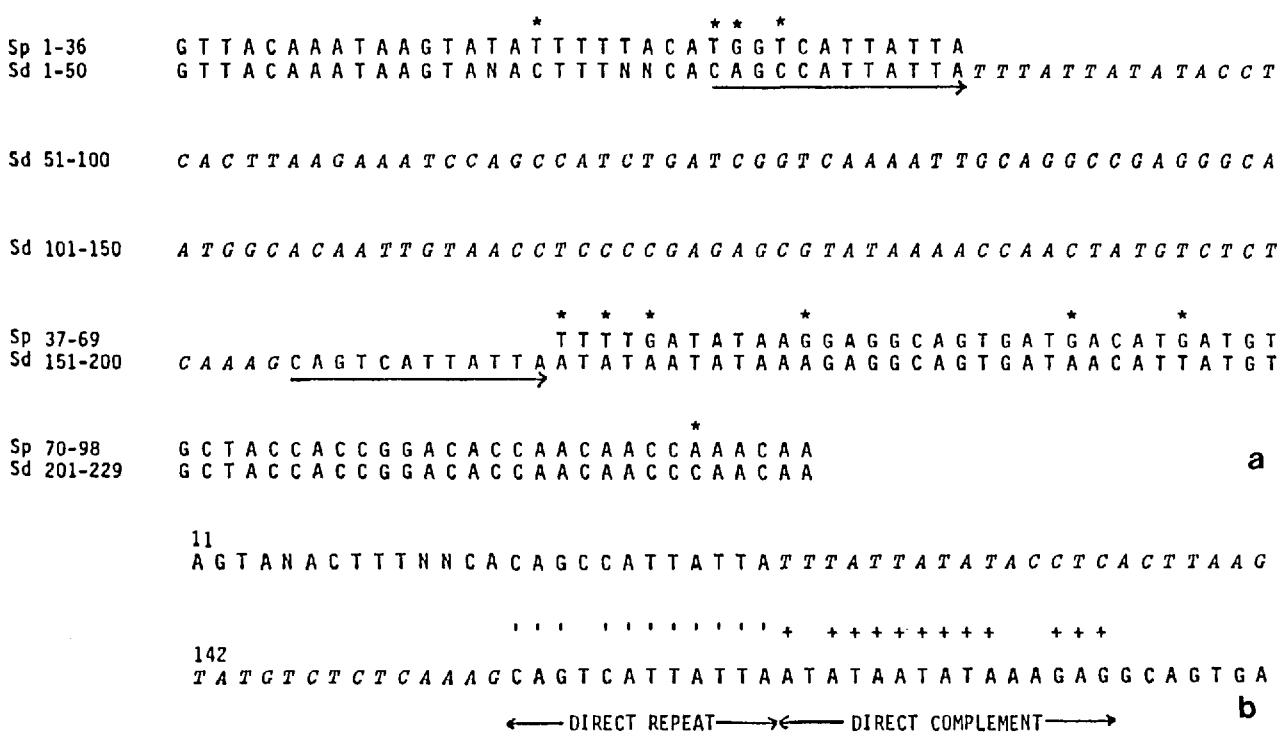

Fig. 3a, b. **a** Nucleotide sequence of the D88I insertion element and comparison of surrounding regions in dRH1.5 and pRH1.4. Regular type represents the resident genomic sequence and italic type represents the D88I element. An asterisk is placed above each mismatch in the sequence. The terminal direct repeats of the D88I element are indicated by an underlining arrow. **b** Sequence homology at the termini of the D88I repeat. This figure shows an interesting feature of the D88I element. Immediately 3′ to the terminal direct repeat in D88I is a 15-NT stretch that has 80% (12/15) direct complement homology with the genomic sequence immediately 3′ to the other direct repeat. This homology is indicated by pluses. Note the lack of any internal inverted repeats, which are characteristic of many mobile repeats. The match-ups for the direct repeats at the D88I termini are indicated by the vertical lines. Only Sd. sequence is shown, and the numbering is the same as in a. D88I sequence is indicated by italicized type, while resident genomic sequence is indicated by regular type. The D88I sequence from nucleotides 59 to 141 has been left out of this figure

The measurement was repeated with Sd. genomic blots, giving a similar result except that each of the "single copy" fragments hybridized to about four bands. It is clear that D88I is a member of a repeated sequence family with many copies in both the Sp. and the Sd. genome, as already indicated by Fig. 2b.

DNA Sequence of the Insertion Elements and the Homologous Sp. Region

Primary DNA sequences were obtained for the region surrounding the insertion site in pRH1.4 and the same region including the D88I element in dRH1.5. Dde/Rsa (dRH1.5) and Dde/Alu (pRH1.4) fragments were fill-in labeled at the Dde1 end and sequenced by the Maxam and Gilbert (1977) method. The sequence of the insertion element and flanking resident sequence is shown in Fig. 3. The inserted sequence is 119 NT long, with a direct repeat of 12 bp (with one mismatch) at the ends. One copy of the direct repeat is present in the Sp. genome at the apparent point of insertion. It agrees in 10 of 12 bases with one of the Sd. repeats and in 9 of 12 with the other.

A genomic DNA sequence apparently was duplicated during insertion, suggesting that D88I is a mobile genetic element. However, examination of the sequence did not reveal any additional direct or inverted repeats such as characteristically occur at the termini of many mobile elements.

An unusual sequence homology exists between the 5′ portion of D88I and what appears to be the previously resident Sd. sequence lying immediately 3′ to the direct repeat. As shown in Fig. 3a, the 5′-most 15 bp of D88I are complementary to the first 15 bp in the flanking sequence immediately 3′ to the direct repeat terminus of the insert in Sd. In Fig. 3b the terminal regions are compared to clarify the nature of the homology. The region of the DNA that was duplicated during the putative insertion event shows clearly as the leftmost 12 nucleotides copied from the Sd. genome. The next 15 nucleotides to the right show the peculiar homology: 12 of 15 bases are complementary to the Sp. sequence. Such a homology is unlikely to occur by chance in this precise

Table 1. Small regions in D88I have significant homology to enhancer core sequences

Sequence name	Alignment
	130
D88I	T A T A A A A C C A A C T A T G T C T
IGCμ	A A T A A A A ˆC C A – C T A G G T A A
	150
D88I	T C A A A G C A G T C A T T A
BK (RC)	C C A A A C C A G T C C T C A
	57
D88I	A A G A A A T C C A G
Py (RC)	T G C A A A A C C A C
	83
D88I	A A T T G C A G G C C G A G G G
Py (RC)	C A T T C C A G G C C T G G G T

A comparison of the D88I sequence with various enhancer core region sequences gave four significant homologous regions. The origin of the enhancer core is given under "Sequence name." The number above the first nucleotide in the sequence comparison indicates the position in the D88I sequence shown in Fig. 3a. Matches are indicated by the dots between the two compared sequences. In each instance, an arrow is drawn below the published core sequence. The viral [human papovavirus BKV and polyoma (Py)] enhancer core sequences are taken from Weiher et al. (1983). The immunoglobulin mu (IGCμ) gene constant region enhancer sequence is from Gillies et al. (1983). "(RC)" indicates that the reverse complement of the published sequence is being used for comparison

location immediately adjacent to the point of insertion. In a random sequence of the same base composition (20% G + C), a match of this accuracy and length would be expected to occur once in 8000 nucleotides.

The ends of the inserted sequence are clearly defined by the 12 nucleotides that are a duplicate of the Sp. sequence at this location. Thus, the 15-NT complementary homology is a relationship between a sequence originally present at the site of insertion and the inserted element D88I. It is not a part of the direct repeat formed at the site of insertion, since it is complementary rather than directly homologous. Conceivably this homology could have been involved in targeting the insertion event. It could also have arisen in a complex rearrangement during the insertion.

Discussion

Evolutionary Divergence

Previous measurements (Hall et al. 1980) showed a median 7% divergence between the single-copy DNA sequences of these two species. Three comparisons between Sp. and Sd. are available for the 88 gene region. Seven of nine restriction sites are held in common in the 7-kb region of overlap between the Sd. and Sp. library inserts. The detailed map of the 1.5-kb region shows that 8 of 17 four-nucleotide sites are held in common. Finally, the region that was sequenced shows 11 substitutions out of 100 nucleotides. These comparisons indicate a divergence not statistically significantly different from the median divergence of the average single-copy DNA sequences.

The Insertion Element

An insertion event has occurred in the transcribed region of the 88 gene of Sd. since the time of the last common ancestor. The following evidence shows that the inserted sequence (D88I) is a member of a family of interspersed repeats with a moderate frequency in Sp., Sd., and Sf. When the Sd. fragment is used as a probe on genome blots of each of the three species, a broad smear of radioactivity is observed (Fig. 2). The fragment containing the repeat (dRH1.5) was used as a probe to screen a lambda library of Sd. (data not presented). At a high criterion (68°C, 0.6 M Na^+), between 0.5% and 1% of the phage showed homology to the repeat, indicating the occurrence of several hundred interspersed members of the D88I family per genome. In a control experiment, the homologous Sp. fragment (pRH1.4) did not hybridize detectably to duplicate transfers of phage.

The D88I element is a candidate for a movable short repetitive sequence with the following characteristics: (a) the observed insertion is bounded by a duplication of a 12-bp preexisting sequence; (b) it has a length of 131 nucleotides; (c) it is homologous to an interspersed family of repeats in each sea urchin species; (d) it lacks inverted terminal repeats or in fact any significant internal repeats other than the site duplication; (e) it lacks a terminal poly(A) region; and (f) it has a 15-NT unusual homology to a sequence immediately adjacent to the point of insertion.

The eukaryotic transposons and suspected transposons that have been examined fall into two major classes: those that contain reverse or direct terminal repeats and those that exhibit a terminal A-rich region [which have been termed retroposons by Lemischka and Sharp (1982)]. As we have just pointed out, D88I does not fall into either of these categories.

D88I is presumably representative of the large class of interspersed repetitive sequences that are known to be inserted frequently into higher-organism DNA on an evolutionary time scale of millions of years (Moore et al. 1978). This work shows that

in echinoderms this process has been occurring for the last few hundred million years and has continued since the existence of the last common ancestor of Sp. and Sd. (5–10 million years ago). The process of insertion is almost certainly continuing at present. Measurements with tracers for total repetitive DNA indicate that the majority of interspersed short repetitive sequences are involved in such processes. While modern recombinant DNA evidence is not available for vertebrate DNA, good arguments indicate that it has had a similar history (Britten 1982).

Two other examples exist in which occurrences of short repeats in the Sp. genome have been sequenced through the terminal regions (Posakony et al. 1983). There is evidence for short insertion-site duplications in each of two occurrences of cs 2109a. These repeats differ from D88I in that they contain internal terminal reverse repeats. Tu1 (Liebermann et al. 1983) is a longer (3.7 kb) sea urchin movable element and thus is not very germane to this discussion, although insertion-site duplications have been observed for several of its occurrences.

Even though the eukaryotic sample is not yet very large, in most cases in which a short repeat has been sequenced, either a few nucleotides of site duplication appear to be present or there is indication that they might have been present but are partially masked by subsequent base substitutions. Thus it appears that insertion processes may often be responsible for the occurrence of short interspersed repeats. The best candidates for the continued insertion of short repetitive sequences into widely spaced locations in eukaryotic genomes are mobile genetic elements acting directly or indirectly. That is, the inserted repeats are or were mobile elements themselves or they were inserted through the action of transposable elements in a process such as levitation (Rubin 1983).

Enhancerlike Sequences in D88I

The sequence of D88I contains elements of homology to viral enhancer sequences that are unlikely to have occurred by chance. We compared D88I with several viral enhancer sequences (from simian virus 40, human papovavirus BKV, polyoma, and Moloney murine sarcoma virus). A region of the immunoglobulin mu gene that shows homology to the viral enhancers was also included in the comparison. Table 1 shows the interesting homologies with D88I which include a region similar to the putative enhancer core (Weiher et al. 1983). At this stage it is difficult to say what the significance of these homologies is. They are unlikely to be chance occurrences; in fact, the 14- out of 16-base homology listed at the top of Table 1 is expected to occur only once in every 4000 kb of random sequence.

While this is not the place for a detailed thoretical proposal, we wish to draw attention to the effects that would occur if parts of the sequence of a mobile repeat could carry out a function similar to that of the viral enhancer sequences. A family of mobile repeats that carried sequences that affected transcription rates in a tissue-specific manner (regardless of their precise location in a gene region) could be an important source of evolutionary variability (Britten and Davidson 1969, 1971). All that would be required would be insertion in either orientation within a few kilobases of a gene region. The probability of affecting transcription of genes might be very high. The selection pressure on such regulatory variation would affect the history of the repeat family as well as the evolution of the species.

Acknowledgments. This work was supported by Program Project Grant GM 20927 from the NIH. S.A.J. was supported in part by a postdoctoral fellowship from the Carnegie Institution of Washington. We also thank Luci Hansen for her skillful secretarial assistance.

References

Anderson DM, Scheller RH, Posakony JW, McAllister L, Trabert SG, Beall C, Britten RJ, Davidson EH (1981) Repetitive sequences of the sea urchin genome. J Mol Biol 145:5–28

Britten RJ (1982) Genomic alterations in evolution. In: Bonner JT (ed) Evolution and development. Springer-Verlag, Berlin, pp 41–64

Britten RJ, Davidson EH (1969) Gene regulation for higher cells: a theory. Science 165:349–358

Britten RJ, Davidson EH (1971) Repetitive and nonrepetitive DNA sequences and a speculation on the origins of evolutionary novelty. Q Rev Biol 46:111–138

Davidson EH, Britten RJ (1979) Regulation of gene expression: possible role of repetitive sequences. Science 204:1052–1059

Gillies SD, Morrison SL, Oi VT, Tonegawa S (1983) A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. Cell 33:717–728

Grula JW, Hall TJ, Hunt JA, Giugni TD, Graham GJ, Davidson EH, Britten RJ (1982) Sea urchin DNA sequence variation and reduced interspecies differences of the less variable DNA sequences. Evolution 36:665–676

Hall TJ, Grula JW, Davidson EH, Britten RJ (1980) Evolution of sea urchin non-repetitive DNA. J Mol Evol 16:95–110

Lemischka I, Sharp PA (1982) The sequences of an expressed rat α-tubulin gene and a pseudogene with an inserted repetitive element. Nature 300:330–335

Liebermann D, Hoffman-Liebermann B, Weinthal J, Childs G, Maxson R, Mauron A, Cohen SN, Kedes L (1983) An unusual transposon with long terminal inverted repeats in the sea urchin *Strongylocentrotus purpuratus*. Nature 306:342–347

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci USA 74:560–564

Moore GP, Scheller RH, Davidson EH, Britten RJ (1978) Evo-

lutionary change in the repetition frequency of sea urchin DNA sequences. Cell 15:649–660

Moore GP, Pearson WR, Davidson EH, Britten RJ (1981) Long and short repeats of the sea urchin DNA and their evolution. Chromosoma 84:19–32

Posakony JW, Scheller RH, Anderson DM, Britten RJ, Davidson EH (1981) Repetitive sequences of the sea urchin genome III. Nucleotide sequences of cloned repeat elements. J Mol Biol 149:41–67

Posakony JW, Flytzanis CN, Britten RJ, Davidson EH (1983) Interspersed sequence organization and developmental representation of cloned poly(A) RNAs from sea urchin eggs. J Mol Biol 167:361–389

Rubin GM (1983) Dispersed repetitive DNAs in *Drosophila*. In: Shapiro JA (ed) Mobile genetic elements. Academic Press, New York, pp 329–361

Thomas TL, Britten RJ, Davidson EH (1982) An interspersed region of the sea urchin genome represented in both maternal poly(A) RNA and embryo nuclear RNA. Dev Biol 94:230–239

Weiher H, König M, Gruss P (1983) Multiple point mutations affecting the simian virus 40 enhancer. Science 219:626–631