Biochem. J. (1986) 234, 1–11 (Printed in Great Britain)

REVIEW ARTICLE

Structure and function of repetitive DNA in eukaryotes*

Norman HARDMAN
Department of Biochemistry, University of Aberdeen, Marischal College, Aberdeen AB9 1AS, Scotland, U.K.

Introduction

Prokaryotes possess relatively small genomes consisting predominantly of low-copy number DNA sequences. The genome sizes of different species vary by less than an order of magnitude (Kingsbury, 1969). In contrast, eukaryotic genomes are generally much larger than their prokaryotic counterparts, and a far greater proportion of this DNA (about 30–40%) is repeated (Britten & Kohne, 1968; Laird, 1971). This repetitive component takes on several guises, and for the purpose of previous discussions it has often been useful to classify these sequences according to their various characteristic properties: structure, distribution and reiteration frequency (Jelinek & Schmid, 1982). Recent applications of DNA cloning methods and more powerful analytical techniques to study specific sequence families has provided a different perspective on the apparent diversity of structure and organization of eukaryotic repetitive DNA. Does this new information provide us with clues as to what the functions of these sequences might be?

An unresolved, and possibly related, question centres on the considerable variation seen in the haploid nuclear DNA contents, or 'C-values', of eukaryotes (Britten & Davidson, 1969). These differences are found in many different phyla encompassing amphibians (Straus, 1971), plants (Rothfels *et al.*, 1966; Rees & Jones, 1967), insects (Keyl, 1965) and rodents (Mazrimas & Hatch, 1972). C-value variation can, on occasion, be especially dramatic; it is sometimes observed in organisms of the same genus that have virtually identical morphology and karyotype. This is the essential element of the so-called 'C-value paradox'; nearly identical species must express about the same number of genes despite significant differences in their C-values. What is the nature of all this extra non-coding DNA? It does not, as one might have anticipated, all correspond to additional repetitive elements; it also includes a considerable increase in the amount of 'single-copy' DNA. Such considerations led some to conceive radical theories of genetic organization in eukaryotes to provide a function for this extra DNA (Callan, 1967; Britten & Davidson, 1969; Thomas, 1971) which, in their simplest forms, have proved to be incorrect. A more recent theory, based on the involvement of dispersed repetitive sequences, envisages that some of these elements may have a nuclear role in selecting which mRNA molecules enter the cell cytoplasm (Davidson & Britten, 1979). A recent study may offer some support for such an idea (Sutcliffe *et al.*, 1984), but others disagree with the interpretation of these data (Owens *et al.*, 1985). In any event, such proposals may provide a function for a minority of the non-coding sequences in eukaryotic genomes, but what of the remainder?

Attention has periodically focused on the persuasively presented alternative argument that the bulk of the 'extra' DNA in eukaryotes might be 'selfish' and have no function (Doolittle & Sapienza, 1980; Orgel & Crick, 1980). This Review will re-examine the idea in the light of recent studies of the structure of specific eukaryotic repetitive DNA elements.

Satellites, minisatellites and tandemly-repeated genes

DNA satellites can be major components of eukaryotic genomes. They have been studied extensively in a wide range of species. Their properties have been the subject of a comprehensive and eloquent review (John & Miklos, 1979) but a brief summary of the main features of DNA satellites is appropriate.

Irrespective of whether satellites are cryptic or readily resolved from genomic DNA by various density gradient methods, they share one common feature; their sequences are tandemly repeated (Southern, 1975). Apart from this unifying property, DNA satellites are incredibly diverse. Studies in *Drosophila* underline the general variability in the amounts of different DNA satellites in closely related species (Gall & Atherton, 1974; Holmquist, 1975). Some species appear to dispense with satellites almost entirely in the soma (Lauth *et al.*, 1976). From a close, critical analysis of the properties of DNA satellites John & Miklos (1979) concluded that no entirely convincing evidence exists for a function of satellite DNA sequences in somatic tissues, although they may have functional roles in the germ line, for example in the regulation of recombination at meiosis (John & Miklos, 1979) or in their association (in *Drosophila melanogaster*) with specific genes linked with some heterochromatic regions, particularly in the sex chromosomes (reviewed in Spradling & Rubin, 1981). In this context Cooke *et al.* (1982), using a specific cloned DNA segment of a human Y-chromosome satellite as a hybridization probe, showed that these sequences are highly methylated in somatic tissues but selectively unmethylated in the germ line. This is opposite to the situation with specific gene sequences that are inactive when methylated in the germ-line, and are undermethylated only when the genes are actively transcribed (Waalwyjk & Flavell, 1978). This may point to a germ-line function for some satellites which correlates with selective hypomethylation of its sequences, but the true significance of this observation is not yet understood.

A number of hypotheses have been put forward to explain the origin and evolution of the tandemly-repeated structure typical of DNA satellites. These include saltatory replication (Fry & Salser, 1977) and a process termed 'random unequal crossing over' (Smith, 1973,

Abbreviations used: LTR, long terminal repeat; ORF, open reading frame; bp, base pair; kb, kilobase (pair).

* Dedicated to the memory of Patricia Hardman.

1976, 1978). The elements of these hypotheses and their relative merits have been discussed elsewhere (John & Miklos, 1979), but unequal crossing over deserves a summary explanation in the context of this Review. Essentially, this mechanism is envisaged to initiate as a result of rare, illegitimate recombination events that occur either between sister chromatids post-replication, or between homologous chromosomes during meiosis. Such events are believed to occur by chance with reasonable frequency in regions of non-repetitive DNA, misalignment and unequal crossover generating one chromatid with a tandemly-duplicated segment and another with a deletion for the same region. Following the establishment of such tandem repeats, additional unequal crossovers would be expected to occur more readily, since they could proceed by homologous recombination between related sequences of the tandem arrays. Smith (1978) argued that the initial recombination event must proceed without the benefit of nucleotide sequence homology, or that the mechanism might take advantage of the presence of short homologous sequences arising by chance, presumably at random. This does not rule out mechanisms which generate satellites by 'directed' recombination using short, highly-recombinogenic sequences, and the recent work of Jeffreys *et al.* (1985) on hypervariable minisatellite regions in human DNA adds a new dimension to the above hypothesis. These studies show that short (about 10 base pair) regions forming part of the repeating unit in hypervariable minisatellite DNA clusters resemble the signal sequence for generalized recombination in *Escherichia coli* (Smith *et al.*, 1981). Hence, similar sequences might be used for related mechanisms in eukaryotes.

The above theory makes the surprising suggestion that such tandemly-repetitive sequence arrays may be the normal, expected consequence of a situation where unequal crossovers are not actually prevented (Smith, 1978). Although it operates independently of selective pressure it is not difficult to see how such a mechanism could be adopted to amplify selected genes which might confer some phenotypic advantage. Tandemly-repeated genes are commonly found in eukaryotic genomes and many examples can be quoted, including genes for 5 S RNA, histones and rRNA. Amplification of dihydrofolate reductase genes in cells treated with methotrexate is an extreme case of the rapid evolution of a tandemly-repeated eukaryotic gene family under conditions of strong selective pressure (Bostock *et al.*, 1979; Kaufman *et al.*, 1979). Ribosomal RNA genes are likewise an example of how such sequence amplification mechanisms could have been adapted for more than one aspect of their function, both in the amplification of the genes themselves, and their spacers, which appear to contain enhancers of rRNA transcription (Reeder & Roan, 1984). One form of nucleolar dominance, where *X. laevis* rDNA genes are transcriptionally dominant over those for *X. borealis* in interspecies hybrids, can be explained by the presence of a larger number of enhancer-containing tandem repeats in the *X. laevis* rDNA spacer compared with that of *X. borealis* (Reeder & Roan, 1984). These, and further considerations regarding the co-evolution of rDNA structure and function, have been elaborated by others in a number of stimulating articles (Dover, 1982; Dover & Flavell, 1984). Further discussion of these aspects is beyond the scope of this Review.

To summarize the major conclusions: DNA satellites are many and varied, and they are a major fraction of the repetitive DNA in some eukaryotes. Plausable mechanisms have been proposed to account for their formation and perpetuation in eukaryotic genomes which do not necessarily depend on such sequences having a structural or functional role, but special cases can be cited where such mechanisms could have provided conceivable selective advantages during the course of eukaryote evolution.

Transposable genetic elements

A large number of discrete genetic elements that are capable of transposition from one genomic location to another have been identified in *Escherichia coli* and other micro-organisms (Kleckner, 1981). The details of the transposition mechanisms have not firmly been established in all cases, but it is generally recognized as being a two-step process occurring at the DNA level, involving co-integration and resolution of the donor and recipient DNA molecules, thus generating a duplicated copy of the transposable element at the target site. It is known that such elements have the capacity to code for the enzymes involved in the transposition process. This is true irrespective of whether they are insertion sequences, devoid of additional genes (Bennett *et al.*, 1980), or transposons, which also possess phenotypic markers that can provide a convenient means to monitor their location (Sherratt *et al.*, 1981). Where such elements integrate into specific genes they generate insertional mutations, and they are responsible for a wide variety of related effects, including deletions and other genetic rearrangements (Nevers & Saedler, 1977; Ghosal & Saedler, 1978). As summarized below, a wide body of evidence suggests that diverse collections of mobile genetic elements similarly exist in eukaryotic genomes.

Retroviruses and retrotransposons. Retroviruses have been a subject of intense interest because of their ability to induce carcinogenesis by transduction of cellular oncogenes (Bishop, 1983). Various classes of retroviruses have been identified, and their genomes all share basically similar structural features. These include genes for the reverse transcriptase/RNAase H involved in virus replication (*pol*), structural components of the virion (*gag* and *env*) and other virus-coded proteins such as a site-specific proteinase and a DNA endonuclease (Weiss *et al.*, 1982; Von der Helm, 1977; Grandegenett *et al.*, 1978). This complement of genes is flanked by long terminal repeated sequences (LTR sequences) which contain essential control elements and other sequences recognized at different stages during the replication/life cycle of the virus (Temin, 1981, 1982). These include a transcriptional start signal, enhancers, a poly(A)-addition signal, and short inverted terminal repeats of a few base pairs, terminated by 5′ TG...CA 3′ (Fig. 1). These small terminal repeats are recognised by the recombination mechanism which leads to integration of the double-stranded retrovirus cDNA into cellular DNA (Majors & Varmus, 1981; Temin, 1981; Chen & Barker, 1984) probably involving the virus-coded endonuclease (Duyk *et al.*, 1983). Short sequences located immediately adjacent to the left and right LTRs in the internal domain of the retrovirus genome are concerned with initiation of (−) and (+) strand DNA synthesis from the viral RNA template (Varmus, 1982). These similarities in structural

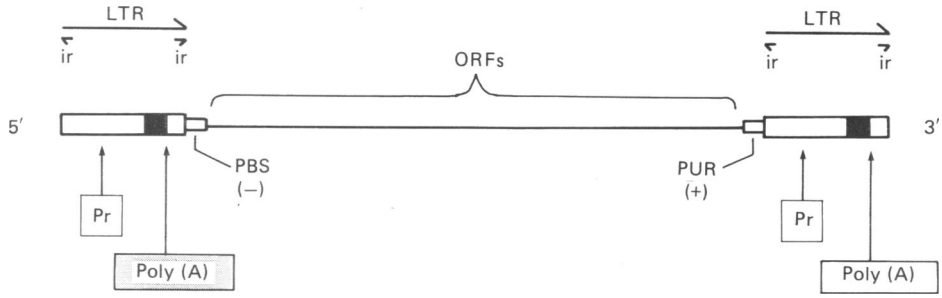

Fig. 1. Schematic generalized diagram of the structure of a retrovirus provirus or a retrotransposon

Long terminal repeated sequences (LTRs) containing transcriptional promoters (Pr), poly(A) additional signals and short terminal inverted repeats (ir), border the element. The internal domain contains structural genes in various arrangements and combinations, depending on the element concerned, specified by open reading frames (ORFs) in the nucleotide sequence. Transcription of these genes is initiated from the 5′ LTR promoter, through the supposedly unutilized 5′ poly(A) addition site, and terminates in the 3′ LTR. PBS and PUR are the tRNA and purine-rich primer binding sites important for reverse transcriptase-directed (−) and (+) strand cDNA synthesis from the RNA transcript. Portions of the LTRs (R-regions, black segments) are probably involved in generating cyclic cDNA intermediates in the generation of new, integrated DNA copies of the element.

organization and mode of virus replication extend to certain other eukaryotic viruses, including cauliflower mosaic virus (Hohn *et al.*, 1985) and hepatitis B virus (Summers & Mason, 1982).

A number of reviews have drawn attention to the striking structural similarities between retroviruses and mobile genetic elements, both in prokaryotes (Kleckner, 1977, 1981) and in eukaryotic genomes (Calos & Miller, 1980; Temin, 1980; Starlinger, 1984; Baltimore, 1985). Many families of eukaryotic mobile genetic elements have been identified, including several in *Drosophila* (Spradling & Rubin, 1981), and an increasing number in the genomes of other organisms, ranging from yeast (Kingsman *et al.*, 1981), slime moulds (Cappello *et al.*, 1984), nematodes (Liao *et al.*, 1983), plants (Shepherd *et al.*, 1984; Johns *et al.*, 1985) and vertebrates (Keshet & Shaul, 1981; Burt *et al.*, 1984). Structural similarities with retrovirus genomes are not limited only to the presence of LTR segments and their short inverted terminal repeats (Kugimiya *et al.*, 1983), but also include predicted amino acid sequence homologies within open reading frames (ORFs) situated within the internal domains of these genetic elements (Fig. 1), corresponding to the retroviral reverse transcriptase, proteinase and endonuclease genes (Saigo *et al.*, 1984; Toh *et al.*, 1985; Hauber *et al.*, 1985). This is in keeping with the hypothesis that these genetic elements may share the same evolutionary origin (Temin, 1980), and that they might replicate/transpose using basically similar mechanisms (Mellor *et al.*, 1985). A simplified scheme illustrating the analogies between the replicative cycles of retroviruses and mobile genetic elements is shown in Fig. 2. This common pathway relies on the idea that mobile genetic elements are analogous to the integrated provirus of retroviruses, and that copies of such elements transpose to new genomic locations by a process which includes transcription, followed by reverse transcription and genomic integration. Defective virus-like particles seen in some eukaryotic systems (Lueders & Kuff, 1977; Burt *et al.*, 1984) suggest that there may exist a complete spectrum of genetic elements, ranging from those that have evolved complete cellular autonomy, such as the retroviruses, to those that might be obligate cellular or genomic parasites, such as the mobile genetic elements. Further analogies derive from the detection of retrovirus-like particles containing RNA transcripts of the *Copia* transposable element (Shiba & Saigo, 1983), and also of extrachromosomal double-stranded circular DNA in *Drosophila* cells, some corresponding to *Copia* DNA (Stansfield & Lengyel, 1979; Flavell & Ish-Horowicz, 1981, 1983). It is thought that these DNA molecules may represent possible transposition intermediates, since similar extrachromosomal DNA structures are found for retroviruses (Shoemaker *et al.*, 1981). This may not be the origin of all extrachromosomal *Copia* DNA sequences, since some could conceivably be excised from the genome and/or be capable of replicating autonomously, as indicated in Fig. 2 (Sinclair *et al.*, 1983). Using genetically 'tagged' copies of the yeast transposable element Ty, Boeke *et al.* (1985) have provided the first direct evidence for the involvement of an RNA intermediate in the transposition process, and have coined the term 'retrotransposon' to describe elements that transpose using a retrovirus-like mechanism. No doubt a similar approach will be used in the near future to investigate the mechanism of transposition of other eukaryotic mobile genetic elements.

Although the retrotransposition mechanism contrasts with that seen in prokaryotes involving DNA intermediates (Kleckner, 1981) both processes are replicative; a copy of the parental element remains at its original location and new copies are generated at the transposition target site. The replicative nature of such transposition processes, combined with the potential of mobile genetic elements to encode for proteins involved in their own transposition, may be the only properties required to ensure their survival as components of prokaryotic and eukaryotic genomes; as with DNA satellites it is not necessary to ascribe a function to such elements as a precondition of their propagation.

As indicated above, bacteria, and some eukaryotes such as yeast, possess genomes with little or no repetitive DNA other than that which can be accounted for by families of transposable genetic elements. The genomes of other eukaryotes such as *Aspergillus* may lack repetitive DNA altogether, apart from multiple rDNA

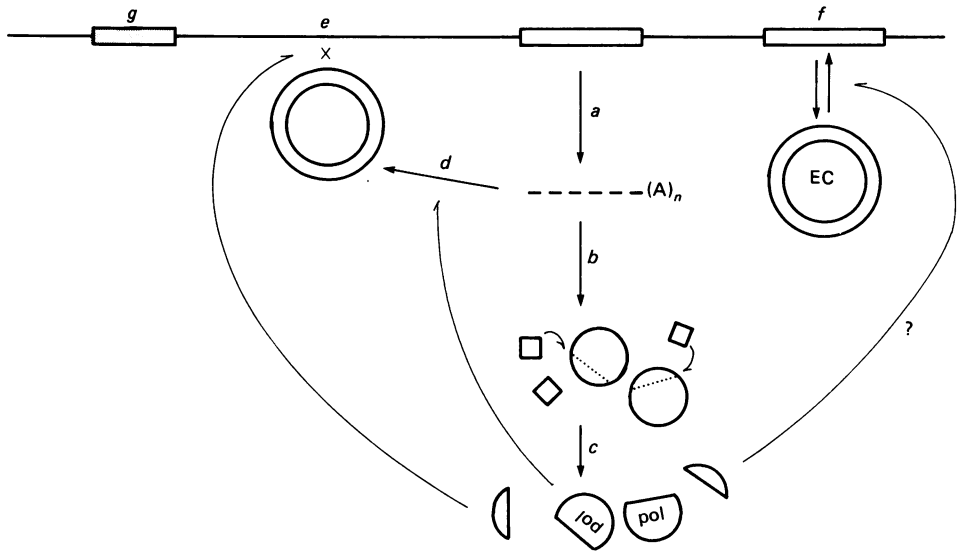

Fig. 2. Replicative cycles of retroviruses and retrotransposons

(*a*) The integrated DNA of the retroviral provirus DNA, or retrotransposon, is transcribed to generate long, polyadenylated RNA transcripts (broken line) containing R-regions of both LTRs. (*b*) RNA transcripts are translated to produce structural proteins (for retroviruses), or a proteinase (squares) involved in processing of other gene products (*c*) involved in transposition, such as reverse transcriptase (pol) or DNA endonuclease (see text). (*d*) The RNA transcript also serves as a template for reverse transcriptase-directed synthesis of circular cDNA intermediates capable of integrating at new genomic locations, possibly with the aid of the retrovirus/retrotransposon-coded endonuclease (*e*). The same gene products might also be used in *trans* (f) in the generation and/or integration of related or of other unrelated extrachromosomal DNA intermediates (EC). Truncated, or otherwise inactive, copies of the element might retain the ability to become integrated but not transcribed (*g*, shaded).

genes (Timberlake, 1978). In eukaryotic genomes containing much larger components of interspersed repetitive DNA it is therefore pertinent to ask what *portion* of this repetitive DNA can be attributed to families of mobile elements?

Properties of 'middle-repetitive' DNA sequences

DNA satellites are generally the most highly-repetitive sequence components in eukaryotic DNA (Britten & Kohne, 1968). 'Middle-repetitive' DNA is a term usually used as a broad description of an additional heterogeneous sequence component consisting of many different families of lower-copy-number repetitive elements which collectively comprise a major fraction (30–40%) of the DNA in most eukaryotic genomes (Britten & Kohne, 1968). Disagreement exists over the level of interspersion of middle-repetitive DNA with other sequence components in some genomes (Manning *et al.*, 1975; Biezunski, 1981*a*, *b*; Moyzis *et al.*, 1981*a*, *b*). Attention will be focused primarily on new studies in a few selected organisms where detailed information has become available on the structure and distribution of specific middle-repetitive elements.

Middle-repetitive elements in *Drosophila* DNA. *Drosophila melanogaster* and related species have proved an especially useful experimental system in which to address the question of what portion of eukaryotic middle-repetitive DNA may consist of mobile genetic elements. Their primary advantage is that the technique of *in situ* hybridization to salivary gland polytene chromosomes can be used to map the location of all the members of a dispersed middle-repetitive sequence family by utilizing a specific, cloned repetitive element as a hybridization probe (Wensink *et al.*, 1974). Second, using the same technique it is possible to compare the chromosomal positions of the same family of repetitive elements within the DNA of individuals within a given fly stock, between different strains of the same species of *Drosophila*, and between *Drosophila melanogaster* and sibling species (e.g. *D. simulans*). A number of striking observations have been made using this approach.

Approx. 12% of the genome of *Drosophila melanogaster* consists of 'middle-repetitive' DNA (Brutlag *et al.*, 1977). About one-quarter of this component consists of dispersed tRNA genes and tandemly-repeated genes coding for histones, rRNA and 5 S RNA (reviewed by Spradling & Rubin, 1981). The remainder consists of about 50 or more families of dispersed repeated elements containing between 10 and 100 sequences per family. Using a panel of seventeen dispersed middle-repetitive DNA sequences selected at random by cloning, Young (1979) showed that the locations of some or all of the sequences differed in the polytene chromosomes of two non-interbreeding strains of *Drosophila melanogaster*, indicating that in all cases the sequences were derived from families of mobile genetic elements. Similar experiments have been performed in several other laboratories (Rubin *et al.*, 1981; Ananiev *et al.*, 1984; Hunt *et al.*, 1984; Junakovic *et al.*, 1984). Some of these sequences correspond to well-characterized families of transposable genetic elements, including *Copia*-like sequences (*Copia*, 412, 297, 17.6, mdg1, mdg3, B104; Rubin *et al.*, 1981; Scherer *et al.*, 1982) and other distinct families of mobile elements including FB elements (Potter *et al.*, 1980), Gypsy (Modolell *et al.*, 1983), P-elements (Rubin *et al.*, 1982), hobo (McGinnis *et al.*, 1983), I-factors (Bucheton *et al.*, 1984) and less well-characterized mobile

elements (Young, 1979). It has been estimated that these families of dispersed transposable genetic elements collectively may total over 30 and account for most of the remaining 75% of the middle-repetitive DNA in *Drosophila melanogaster* and related species (Spradling & Rubin, 1981). The locations of these dispersed mobile elements are generally conserved within an inbred fly population (Ananiev *et al.*, 1984) and variant between separate stocks of the same species (Young, 1979; Junakovic *et al.*, 1984). Moreover, some families of transposable elements may be absent altogether from closely related species of *Drosophila* (Dowsett & Young, 1982; Hunt *et al.*, 1984). The remainder of the middle-repetitive elements appear to be confined to constant positions at specific chromosomal locations, including the pericentromeric regions of polytene chromosomes (Dowsett & Young, 1982). Recent careful studies (Ananiev *et al.*, 1984) have revealed several other significant findings concerning the properties of mobile dispersed middle-repetitive elements in *Drosophila*. These include the observation that some families of elements may 'prefer' to transpose into similar genomic locations; that the presence of a number of such elements at a single chromosomal region does not affect chromosome morphology; that polytene bands with the largest DNA contents probably offer the largest targets for transposition; that the regions of DNA surrounding centromers may be composed almost entirely of clusters of mobile elements.

From this large amount of structural information it is possible to come to several important conclusions: (1) the majority of middle-repetitive DNA in *Drosophila* consists of potentially mobile genetic elements; (2) the chromosomal location and copy number of a given mobile middle-repetitive element is under close genetic control within a given fly population, and (3) most of the dispersed middle-repetitive DNA provides no function essential to the survival of these insects. Are such observations a peculiarity of insects? What of the middle-repetitive DNA in other higher organisms?

L1 elements and pseudogenes in mammalian DNA. It is not necessary to present a totally comprehensive account of the properties of repetitive DNA in mammals; this subject has recently been reviewed elsewhere (Jelinek & Schmid, 1982). Instead, this section will concentrate on the two most abundant and best-characterized middle-repetitive sequence families in mammalian DNA: the short, interspersed *Alu*1 repeats (Houck *et al.*, 1979) and the long, interspersed repeated elements referred to as LINE or L1 elements (Voliva *et al.*, 1983; Singer, 1982; Singer & Skowronski, 1985).

Alu-repeats. The majority of the middle-repetitive DNA in mammalian genomes consists of numerous families of repeated sequences that are only a few hundred base pairs in length (Schmid & Deininger, 1975). One family of short, interspersed repeated elements dominates this reiterated DNA fraction. This sequence family was first described as a component of human DNA, and was called the '*Alu*1-family' since most of its members contain *Alu*1 restriction sites (Houck *et al.*, 1979). There may be up to 500000 copies of *Alu*-repeats in human DNA, accounting for several percent of the genome. *Alu*-equivalent sequences have been reported in other primates as well as in rodents (Grimaldi *et al.*, 1981; Haynes *et al.*, 1981) and they are probably present in

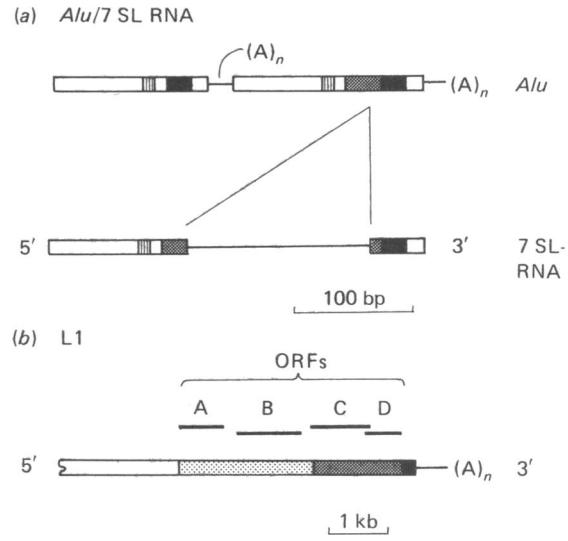

Fig. 3. Structure of *Alu*1-repeats and L1 elements in human DNA

(*a*) The human *Alu*1-repeat: the structure of the *Alu*1-repeat is a dimer of two related tandem repeats, each terminated by an A-rich segment [$(A)_n$]. It is related to the sequence of 7 SL-RNA as indicated. A segment of 155 base residues from the centre of the 7 SL-RNA sequence is absent from the *Alu*1-repeat unit. (*b*) The human L1 repeat: the sequence is given the arbitrary arrangement 5′–3′. The 5′ ends of different L1 elements are truncated. The 3′ ends are conserved and terminated by A-rich sequences [$(A)_n$]. Regions of the sequence having potential open reading frames (ORFs) A–D, which have the same 5′–3′ orientation, are indicated. Regions showing sequence and hybridization homology with the rodent L1 homologue are shown as darkly- and lightly-shaded segments, respectively.

abundance in *Xenopus* DNA (Ullu & Tschudi, 1984). *Alu*1 repeats in human DNA usually consist of a head-to-tail tandem arrangement of two related sequences about 130 bp long, each terminated by a A-rich segment. This is shown schematically in Fig. 3(*a*). One of the sequences contains an additional, internal segment of 32 bp (Deininger *et al.*, 1981). The equivalent sequence in rodents is derived from just one 130 bp repeating unit, containing an unhyphenated tandem repeat formed by duplication of an internal 30 bp sequence (Kalb *et al.*, 1983).

Recent studies have revealed highly-significant (about 80%) sequence homology between the longer unit of the human *Alu*1 consensus sequence and the 5′ and 3′ portions of 7 SL RNA, the abundant cytoplasmic RNA, 300 bases in length, which forms part of the signal recognition particle (Walter & Blobel, 1980). As summarized in Fig. 3(*a*), the central 155 bp of the 7 SL RNA primary sequence is not represented in the *Alu*1-repeat (Ullu & Tschudi, 1984). This interesting work has provided an important insight into the evolution of *Alu*1-repeats in mammalian DNA. No *Alu*-equivalent repeats, and only two genes coding for 7 SL RNA, are found in the *Drosophila* genome (Gundelfinger *et al.*, 1984). Analysis of 7 SL RNA from man, *Xenopus* and *Drosophila* indicates that its sequence is subject to strong evolutionary conservation (Ullu & Tschudi, 1984). Taken together these data indicate that 7 SL RNA is the

progenitor of the *Alu*1-sequence family, and that the evolutionary process which led to the generation and dispersal of *Alu*1-repeats preceded mammalian radiation. The absence of *Alu*1-repeats and similar elements in *Drosophila* DNA may help to explain the lack of a short-period interspersion pattern of small middle-repetitive sequences in this organism (Manning *et al.*, 1975).

Evidence points to the involvement of an RNA intermediate in the evolution of *Alu*1-repeats, at least in the early stages. Limited digestion of 7 SL RNA in the intact signal recognition particle generates three fragments, two corresponding in size to the 5′ and 3′ *Alu*1-repeat related segments and the third to the internal 7 SL RNA-specific portion of the molecule (Gundelfinger *et al.*, 1983). It can thus be argued that *Alu*1-repeats are processed 7 SL RNA transcripts, containing 3′ poly(A) segments which may have provided the template for their reverse transcription into cDNA prior to genomic integration. In keeping with this hypothesis there are striking structural analogies between *Alu*1-repeats and processed pseudogene copies of other defined gene products, including the small nuclear RNAs (snRNAs) and mRNAs for globin, immunoglobulins, tubulins and metallothionein (reviewed by Sharpe, 1983). In no case has the mechanism leading to the generation of pseudogenes been totally resolved, although it is interesting to note that short direct repeats flank both *Alu*1-repeats and pseudogenes, indicative of a common process which might involve a retrovirus transposase-like enzymic activity.

L1 elements. L1 elements are believed to be the only major family of long, interspersed repeated elements in primate DNA. Homologues exist in the genomes of rodents and probably other mammals. There are $(1–4) \times 10^4$ copies of these elements in the human genome. They vary in size up to 6–7 kb and account for at least 2–3% of the total DNA complement (Singer, 1982). Different segments of the rodent homologue were cloned independently as separate sequences (Fanning, 1982; Gebhard *et al.*, 1982), but were later shown to be colinear (Fanning, 1983; Bennett & Hastie, 1984). The properties of L1 elements have been reviewed recently (Rogers, 1984; Singer & Skowronski, 1985) so their salient features will be summarized briefly.

Cloned human and rodent L1 elements are heterogeneous in length. Smaller versions are truncated at one end (designated 5′) but they contain the same 3′ sequences terminated by an A-rich segment of variable length (Lerman *et al.*, 1983; Grimaldi *et al.*, 1984) which corresponds to the 3′ end of RNA transcripts *in vivo* (DiGiovanni *et al.*, 1983). These properties, as in the case of *Alu*1-repeats, have been compared with those of processed pseudogenes for mRNAs and snRNAs (Jagadeeswaran *et al.*, 1981; Van Arsdell *et al.*, 1981) and suggest that L1 elements may also have been dispersed in a fashion analogous to retrovirus proviruses. Short sequence duplications bordering individual L1 elements can be identified, consistent with a mechanism which generates staggered DNA strand breaks at the point of their insertion (Grimaldi *et al.*, 1984; Rogers, 1984, 1985). Longer elements of the rodent and human families are 60% homologous for over 1500 bp near their 3′ ends, and display hybridization cross-homology for a further 2000 bp encompassing four long open-reading frames of over 600 bp, all oriented in the 5′–3′ direction (Voliva *et al.*, 1983). These data, summarized in Fig. 3(*b*), indicate that the L1 family of dispersed repeated sequences consists largely of incomplete, probably non-functional, processed pseudogene-like copies together with a small number of transcriptionally-active elements (Martin *et al.*, 1984; Rogers, 1985; Singer & Skowronski, 1985). In contrast with retrotransposons, however, the L1 elements so far studied are devoid of LTR sequences, although it cannot be excluded that the actively-transcribed progenitor elements do not possess them since, in all probability, none have yet been isolated (Rogers, 1985).

Recent studies have shed new light on the developmental stage at which the putative progenitor L1 elements may be transcriptionally active. It has long been known that they form part of RNA polymerase II-directed transcripts of heterogeneous nuclear RNA in a number of somatic cell lines (Lerman *et al.*, 1983; Shafit-Zagardo *et al.*, 1983). The heterogeneous size of these transcripts, their confinement to the nucleus, and the fact that they contain sequences other than L1 elements, points to the conclusion that the copies of L1-repeats that they contain are transcribed adventitiously as part of other RNA molecules because of their wide genomic distribution. In contrast, an abundant specific 6.5 kb-long cytoplasmic polyadenylated L1 transcript, corresponding to the coding strand for the open reading frames A–D shown in Fig. 3(*b*), is detected only in undifferentiated, and not in differentiated, human teratocarcinoma cells (Skowronski & Singer, 1985). Brulet *et al.* (1983) similarly showed that a family of mouse retrovirus-like sequences are selectively transcribed in embryonic cells in a manner possibly analogous to the programmed transcription of *Copia*-like elements that occurs early in *Drosophila* development (Spradling & Rubin, 1981). Further similarities between L1 elements and retrotransposon sequences includes their association with DNA rearrangements (Lerman *et al.*, 1983; DiGiovanni *et al.*, 1983) and their occasional presence as extrachromosomal DNA species (Schindler & Rush, 1985).

What can be concluded from these recent studies of *Alu*1-repeats and L1 elements? These sequences appear to be mobile, like the middle-repetitive DNA elements of *Drosophila*. Collectively they account for about 6–8% of the human genome, or about 20% of the middle-repetitive DNA fraction. Structurally they differ from the retrotransposon-like elements which account for the bulk of the dispersed middle-repetitive DNA in *Drosophila*, but the available evidence points to a similar mechanism for their dispersal involving retrotranscription. Like other pseudogenes, *Alu*1-repeats seemingly have arisen from RNA transcripts of a defined gene and they undoubtedly do not encode proteins required for transposition; presumably they rely for their dispersal on gene products provided in *trans* from other genetic elements. This may not be the case for the putative progenitor L1 elements, since they have some properties in common with retrotransposon-like middle-repetitive sequences in *Drosophila*: they possess potential internal open reading frames which may code for transposition functions, and evidence points to their specific expression early in development. Many families of pseudogenes and transposon-like elements have already been characterized and it is a reasonable speculation that new families will be discovered, accounting for a significant additional fraction of mammalian middle-repetitive DNA.

Foldback sequences and clustered repetitive elements. 'Foldback DNA' is a term originally coined by Wilson & Thomas (1974) to describe the snap-back DNA structures formed when eukaryotic DNA is denatured and allowed to anneal at low DNA concentrations to avoid intermolecular reassociation. They result from the presence of inverted repeat sequences located within the same DNA fragment, and account for a variable though significant fraction (1–10%) of the DNA in most eukaryotic genomes. The general properties and distribution of foldback sequences have been studied in a wide range of eukaryotes, from slime moulds to mammals (Cech & Hearst, 1975; Schmid *et al.*, 1975; Deininger & Schmid, 1976; Hardman & Jack, 1977). In this early work much attention was paid to studying differences in the distribution of foldback elements in different species and making correlations between the properties of foldback DNA and of middle-repetitive DNA sequences (Schmid *et al.*, 1975; Hardman *et al.*, 1979*b*, 1980). Almost a decade ago a reassociation-kinetic study of total *Xenopus laevis* foldback DNA led to the suggestion that these sequences may be mobile genetic elements (Perlman *et al.*, 1976). Except in the case of the eukaryotic slime-mould *Physarum polycephalum* no detailed studies using specific, cloned foldback elements have been carried out to investigate this assertion further.

Foldback sequences are widely distributed in the genome of *Physarum polycephalum* (Peoples *et al.*, 1983). Foldback DNA structures formed in large (30–100 kb) *Physarum* nuclear DNA fragments fall into three categories: structures formed from the extrachromosomal nucleolar rDNA satellite, small dispersed foldback sequences and complex foldback structures generated from clusters of generally longer inverted repetitive DNA sequences (Hardman *et al.*, 1979*b*). Similar observations have been made using mammalian DNA (Hardman *et al.*, 1979*a*; Biezunski, 1981*b*) and *Drosophila* DNA (Biezunski, 1981*a*). The complex, clustered foldback DNA structures are of special interest in the context of this Review since they account for over one-half of the mass of foldback DNA in *Physarum*, and are generated from genomic DNA segments over 20–50 kb in length occupied almost exclusively by scrambled clusters of a single family of long, methylated, highly-repetitive sequences (Peoples & Hardman, 1983). These elements have been referred to as *Hpa*II-repeats, and they account for one-half of the middle-repetitive DNA fraction and up to 15–20% of the nuclear DNA complement in *Physarum* (Peoples *et al.*, 1985). Nucleotide sequence analysis of cloned *Physarum* DNA segments containing *Hpa*II repeats indicate that this repetitive element is about 8.6 kb in length and contains directly-repeated, LTR-like termini having structural features in common with the *Drosophila* transposable element *Copia* (Pearston *et al.*, 1985). It is suspected that clusters of *Hpa*II repeats result from transposition of these elements in either orientation into target sites located in other copies of the same sequence, leading to sequence scrambling and the generation of DNA segments with the capacity to form a variety of complex foldback DNA structures. Targeted insertion of the transposable element DIRS-1 into its own sequences has similarly been observed in *Dictyostelium discoideum* (Cappello *et al.*, 1984). An interesting consequence of such targeted insertions is the interruption of resident, target transposable elements and, presumably, their inactivation. Thus, most loci containing these elements may be non-functional scrambled remnants of previous transposition events that have become fixed in the genome.

Scrambled clusters of repetitive elements are not confined to the genomes of the slime moulds; they have also been observed in the DNA of *Drosophila melanogaster* (Wensink *et al.*, 1979), chicken (Eden *et al.*, 1981; Sobieski & Eden, 1981), hamster (Hardman *et al.*, 1979*a*; Moyzis *et al.*, 1981*a*, *b*) and mouse (Biezunski, 1981*b*). As in *Physarum*, clustered repeats may account for a substantial fraction, possibly up to one-third in some cases (Eden *et al.*, 1981), of the repetitive DNA fraction, but it is not known with certainty in other genomes how many sequence families are involved in such scrambled clusters or if they result from transposition-like events involving nomadic elements. However, it is not difficult to envisage how scrambled clusters of transposon-like elements could readily arise as a result of multiple insertions at locations where there was little or no phenotypic selection against their accumulation. Indeed, in *Drosophila* DNA, immobile clusters of middle-repetitive sequences are present in the pericentromeric regions of chromosomes and in certain other locations known to be preferred transposition target sites (Dowsett & Young, 1982; Ananiev *et al.*, 1984), but the nature and arrangement of these sequences remains to be elucidated.

Clearly, many questions remain concerning the origin of foldback DNA sequences and the more complex repetitive sequence arrangements seen in some eukaryotic genomes. However, as indicated above, at least one abundant sequence family capable of forming clustered foldback sequences has all the hallmarks of a transposable element. It is therefore unnecessary to invoke any fundamentally new principles to explain the origin of foldback structures and repetitive sequence clusters in eukaryotic DNA; they may arise as a natural consequence of the spread of repetitive DNA sequences by processes such as DNA transposition.

The C-value problem

If potentially selfish sequences such as mobile genetic elements and DNA satellites have the capacity to accumulate in eukaryotic genomes, could differences in the proliferation of these sequences offer a plausible solution to the problem of C-value variation? When addressing this question it is important to consider the different controls which may serve to limit the spread of selfish sequences and thus account for the large C-value variations that are observed. Here we are mostly confined to general ideas and speculation rather than critical analysis of data.

The work of Bennett (1971) and Olmo & Morescalchi (1975) suggests that a positive correlation exists between the DNA content per nucleus (in plants and amphibians) and several growth parameters that could be subject to the influences of natural selection, including cell size and mitotic cycle time (reviewed by John & Miklos, 1979). Hence limitations on genome size may, under some circumstances, provide a crude, general mechanism for controlling the overal complement of selfish DNA (Koch, 1972). As pointed out by Doolittle & Sapienza (1980) the presence of more than a small amount of non-functional parasitic DNA may be an intolerable energetic burden for rapidly-dividing organisms such as prokaryotes and primitive eukaryotes with comparatively small, streamlined genomes. This would provide strong selective

pressure in favour of stringent control of the proliferation of non-functional sequences, and their elimination using the mechanisms of genetic recombination. Some evidence in support of this idea has arisen unexpectedly from studies that have involved cloning eukaryotic DNA sequences in bacterial cells. It is well known that recombination-deficient *recA*$^-$ strains of *Escherichia coli* are required for the stable propagation of many cloned eukaryotic DNA sequences. Even in *recA*$^-$ hosts, 'random' segments of *Physarum* DNA, for example, progressively lose internal sequences after cloning which map to the position of repetitive 'foldback' elements (Peoples *et al.*, 1983). It is possible to minimise the loss of such sequences by using *recBC*$^-$/*sbcB*$^-$ hosts, defective in other steps of the recombination pathway (Leach & Stahl, 1983; Nader *et al.*, 1985; Wyman *et al.*, 1985). This information is not only of practical value when attempting to generate representative eukaryotic DNA libraries, but also serves to demonstrate that some dispersed eukaryotic repetitive DNA sequences are rapidly and effectively eliminated from bacterial cell replicons, where different selective pressures are operative, unless elaborate steps are taken to disable the recombination pathway of the host.

The wide disparity in the copy number of different transposable element families, from a few in some cases (Johns *et al.*, 1985) to several thousand in others (Pearston *et al.*, 1985) indicates that additional, more selective factors may also control the spread of an individual family of mobile elements. This could be due to accumulated effects, arising from the various steps in the retrotransposition pathway by which such elements proliferate (Fig. 2), leading to differences in the efficiency with which the 'progenitor' elements are transcribed and their genes expressed in the germ line at appropriate times. Only now are we in a position to begin to study the factors which may influence the transcriptional activity of transposable elements early in development (Brulet *et al.*, 1983; Skowronski & Singer, 1985). Apart from factors affecting their spread, mobile elements can be removed by excision (Dowsett & Young, 1982; Ananiev *et al.*, 1984), in some cases with great precision (O'Hare & Rubin, 1983). These processes are likely to be affected by additional determinants, such as the DNA sequence of the termini of the elements themselves, possibly the nature of the transposition target sites, and the availability and specificity of the enzymes involved in excision.

The above, mostly hypothetical, arguments might account for large fluctuations in the type and quantity of repetitive DNA sequences, but what of the single-copy DNA component which also contributes significantly to C-value variation? Some clues can be found by considering the evolution of intervening sequences in eukaryotic genes.

Gene introns. It would have been surprising if no phenotypic effects resulted from the insertion of mobile genetic elements into eukaryotic genes, and indeed many are seen which parallel the observations made in prokaryotes (McLintock, 1956; Green, 1980; Bingham & Judd, 1981; Levis & Rubin, 1982, McGinnis *et al.*, 1983; Perlman, 1983). These include examples of non-lethal transposable element insertions into defined genetic loci, both in plants (Johns *et al.*, 1985) and in *Drosophila* (Zachar & Bingham, 1982) where they may be a significant source of natural mutations. Non-lethal gene insertions which have no severe phenotypic effect can clearly be tolerated.

Such considerations can be used as an explanation of the origin of gene introns. Mobile elements might be expected to accumulate in genes over prolonged periods of evolutionary time, especially if their effects could be negated by the evolution of enzymic or self-splicing mechanisms (Kruger *et al.*, 1982) capable of precisely excising the intervening sequences during mRNA processing, thus restoring the integrity of the gene product. If such insertions became fixed, their sequences would diverge as part of the gene and gradually be absorbed into the single-copy DNA sequence complement. Introns in eukaryotic DNA can expand perhaps more than 20–30-fold the amount of DNA occupied by a given genetic locus (Nunberg *et al.*, 1980) thus accounting for a considerable proportion of the non-coding single-copy DNA complement. An unknown, but possibly significant, additional fraction of the single-copy DNA complement in other non-coding regions of eukaryotic genomes might have evolved in this way. Unfortunately, it is difficult to imagine how the hypothesis, concerning the origin of the majority of single-copy DNA in eukaryotes, could be tested experimentally.

The question of whether the insertion of mobile elements gave rise to all eukaryotic gene introns remains controversial (Cavalier-Smith, 1985; Cornish-Bowden, 1985). Critics of the idea make the assumption that mechanisms required to remove intervening sequences from gene transcripts must have evolved before mobile elements were permitted to invade genes, and therefore that some introns must have performed some function at an early stage in the evolution of certain genetic loci (Cornish-Bowden, 1985). Such a function might be in allowing exon shuffling and shaping the structural domains of large and complex proteins (Lonberg & Gilbert, 1985; Palm *et al.*, 1985; Stone *et al.*, 1985). Although the selective advantages of such processes cannot be denied, the discovery that some classes of introns can self-splice (Kruger *et al.*, 1982), and that other introns carry genes that, when transcribed, encode 'maturase' proteins involved in RNA splicing (Macreadie *et al.*, 1985) suggests that introns themselves might have provided the information necessary to permit their excision from RNA transcripts of genes into which they inserted.

Whether or not all introns are derived from mobile element insertions will continue to be a source of debate. However, the properties of some classes of introns indicate that they are transposable, and it is likely that they could have inserted into some genes with little or no phenotypic effect. In other cases the position of such insertions could have conferred some selective advantage in the evolution of the gene. In many instances it may be difficult or impossible to distinguish between these possibilities.

Conclusions

Orgel & Crick (1980) defined 'selfish DNA' as having two properties: (1) it arises when a DNA sequence spreads by forming additional copies of itself within a genome, and (2) it makes no specific contribution to the phenotype. At the time that article was written it was already recognized that most satellite DNA sequences are essentially selfish, based on these criteria (John & Miklos,

1979). Although Doolittle & Sapienza (1980) surmised that this concept could be extended to other components of the eukaryotic genome capable of self-propagation, such as the nomadic elements in *Drosophila* DNA, detailed structural information was then known about too few families of middle-repetitive DNA in this, or other, species to be certain of their origin. Subsequent work suggests that this principle may encompass a wide range of transposon-like middle-repetitive DNA families in different species and at least some gene introns. Some sequence families (e.g. the *Alu*1-repeats and pseudogenes) may have originated from RNA transcripts of 'immobile' genes by utilising retrotransposition mechanisms provided *in trans* by mobile DNA elements; their efficiency to invade the genome may partly depend on trivial factors such as the abundance of their transcripts in the germ-line.

Finally, although a number of instances have been cited where such sequences may have provided some phenotypic advantage, as argued previously (Doolittle & Sapienza, 1980; Orgel & Crick, 1980) it may be futile to seek functions for the majority of these potentially 'selfish' eukaryotic repetitive DNA sequences, since none may exist.

I am grateful to colleagues, especially Doug Pearston and Lee Gill, for advice and critical reading of this manuscript, and to Maxine Singer for communicating recent results prior to publication. Work in this laboratory is supported by the Medical Research Council and The Science and Engineering Research Council.

REFERENCES

Ananiev, E. V., Barsky, V. E., Ilyin, Yu, V. & Ryzic, M. V. (1984) Chromosoma **90**, 366–377
Baltimore, D. (1985) Cell **40**, 481–482
Bennett, K. L. & Hastie, N. D. (1984) EMBO J. **3**, 467–472
Bennett, M. D. (1971) Proc. R. Soc. London Ser. B **178**, 277–299
Bennett, P. M., Richmond, M. H. & Petrochulon, V. (1980) Plasmid **3**, 135–149
Biezunski, N. (1981*a*) Chromosoma **84**, 87–109
Biezunski, N. (1981*b*) Chromosoma **84**, 111–129
Bingham, P. M. & Judd, B. H. (1981) Cell **25**, 705–711
Bishop, J. M. (1983) Annu. Rev. Biochem. **52**, 301–354
Boeke, J. D., Garfinkel, D. J., Styles, C. A. & Fink, G. R. (1985) Cell **40**, 491–500
Bostock, C. J., Clark, E. M., Harding, N. G. L., Mounts, P. M., Tyler-Smith, C., van Heyningen, V. & Walker, P. M. B. (1979) Chromosoma **74**, 153–177
Britten, R. J. & Davidson, E. H. (1969) Science **165**, 349–357
Britten, R. J. & Kohne, D. E. (1968) Science **161**, 529–540
Brulet, P., Kaghad, M., Xu, Y.-S., Croissant, O. & Jabob, F. (1983) Proc. Natl. Acad. Sci. U.S.A. **80**, 5641–5645
Brutlag, D., Appels, R., Dennis, E. S. & Peacock, W. J. (1977) J. Mol. Biol. **112**, 31–47
Bucheton, A., Paro, R., Sang, H. M., Pelisson, A. & Finnegan, D. J. (1984) Cell **38**, 153–163
Burt, D. W., Reith, A. D. & Brammar, W. J. (1984) Nucleic Acids Res. **12**, 8579–8593
Callan, H. G. (1967) J. Cell. Sci. **2**, 1–7
Calos, M. P. & Miller, J. H. (1980) Cell **20**, 579–595
Capello, J., Cohen, S. M. & Lodish, H. F. (1984) Mol. Cell Biol. **4**, 2207–2213
Cavalier-Smith, T. (1985) Nature (London) **315**, 283–284
Cech, T. R. & Hearst, J. E. (1975) Cell **5**, 429–446
Chen, H. R. & Barker, W. C. (1984) Nucleic Acids Res. **12**, 1767–1776
Cooke, H. J., Schmidtke, J. & Gosden, J. R. (1982) Chromosoma **87**, 491–502
Cornish-Bowden, A. (1985) Nature (London) **313**, 434–435
Davidson, E. H. & Britten, R. J. (1979) Science **204**, 1052–1059
Deininger, P. L. & Schmid, C. W. (1976) J. Mol. Biol. **106**, 773–790
Deininger, P. L., Jolly, D. J., Rubin, C. M., Friedmann, T. & Schmid, C. W. (1981) J. Mol. Biol. **151**, 17–33
DiGiovanni, L., Haynes, S. R., Misra, R. & Jelinek, W. R. (1983) Proc. Natl. Acad, Sci. U.S.A. **80**, 6533–6537
Dowsett, A. P. & Young, M. W. (1982) Proc. Natl. Acad. Sci. U.S.A. **79**, 4570–4574
Doolittle, W. F. & Sapienza, C. (1980) Nature (London) **284**, 601–603
Dover, G. A. (1982) Nature (London) **299**, 111–117
Dover, G. A. & Flavell, R. B. (1984) Cell **38**, 622–623
Duyk, G., Leiss, J., Longiaru, M. & Skalka, A. M. (1983) Proc. Natl. Acad. Sci. U.S.A. **80**, 6745–6749
Eden, F. C., Musti, A. M. & Sobieski, D. A. (1981) J. Mol. Biol. **148**, 129–151
Fanning, T. G. (1982) Nucleic Acids Res. **10**, 5003–5013
Fanning, T. G. (1983) Nucleic Acids Res. **11**, 5073–5091
Flavell, A. J. & Ish-Horowicz, D. (1981) Nature (London) **292**, 591–595
Flavell, A. J. & Ish-Horowicz, D. (1983) Cell **34**, 415–419
Fry, K. & Salser, W. (1977) Cell **12**, 1069–1089
Gall, J. G. & Atherton, D. D. (1974) J. Mol. Biol. **85**, 633–664
Gebhard, W., Meitinger, T., Hochtl, J. & Zachau, H. G. (1982) J. Mol. Biol. **157**, 453–471
Ghosal, D. & Saedler, H. (1978) Nature (London) **275**, 611–617
Grandegenett, D. P., Vora, A. C. & Schiff, R. D. (1978) Virology **89**, 119–126
Green, M. M. (1980) Annu. Rev. Genet. **14**, 109–120
Grimaldi, G., Queen, C. & Singer, M. F. (1981) Nucleic Acids Res. **9**, 5553–5568
Grimaldi, G., Skowronski, J. & Singer, M. F. (1984) EMBO J. **3**, 1753–1759
Gundelfinger, E. D., Krause, E., Melli, M. & Dobberstein, B. (1983) Nucleic Acids Res. **11**, 7363–7374
Gundelfinger, E. D., di Carlo, M., Zoff, D. & Melli, M. (1984) EMBO J. **3**, 2325–2335
Hardman, N. & Jack, P. L. (1977) Eur. J. Biochem. **74**, 275–283
Hardman, N., Bell, A. J. & McLachlan, A. (1979*a*) Biochim. Biophys. Acta **564**, 372–389
Hardman, N., Jack, P. L., Brown, A. J. P. & McLachlan, A. (1979*b*) Eur. J. Biochem. **94**, 179–187
Hardman, N., Jack, P. L., Fergie, R. C. & Gerrie, L. M. (1980) Eur. J. Biochem. **103**, 247–257
Hauber, J., Nelbock-Hochstetter, P. & Feldmann, H. (1985) Nucleic Acids Res. **13**, 2745–2758
Haynes, S. R., Toomey, T. P., Leinwand, L. & Jelinek, W. R. (1981) J. Mol. Cell Biol. **1**, 573–583
Hohn, T., Hohn, B. & Pfeiffer, P. (1985) Trends Biochem. Sci. **10**, 205–209
Holmquist, G. (1975) Nature (London) **257**, 503–505
Houck, C. M., Rinehart, F. P. & Schmid, C. W. (1979) J. Mol. Biol. **132**, 289–306
Hunt, J. A., Bishop, A. J. G., III & Carson, H. L. (1984) Proc. Natl. Acad. Sci. U.S.A. **81**, 7146–7150
Jagadeeswaran, P., Forget, B. G. & Weissman, S. M. (1981) Cell **26**, 141–142
Jeffreys, A. J., Wilson, V. & Thein, S. L. (1985) Nature (London) **314**, 67–73
Jelinek, W. R. & Schmid, C. W. (1982) Annu. Rev. Biochem. **51**, 813–844
John, B. & Miklos, G. L. B. (1979) Int. Rev. Cytol **58**, 1–114
Johns, M. A., Mottinger, J. & Freeling, M. (1985) EMBO J. **4**, 1093–1102
Junakovic, N., Caneva, R. & Ballario, P. (1984) Chromosoma **90**, 378–382
Kalb, V. F., Glasser, S., King, D. & Lingrel, J. B. (1983) Nucleic Acids Res. **11**, 2177–2184
Kaufman, R. J., Brown, P. C. & Schimke, R. T. (1979) Proc. Natl. Acad. Sci. U.S.A. **76**, 5669–5673
Keshet, E. & Shaul, Y. (1981) Nature (London) **289**, 83–85

Keyl, H. G. (1965) Experientia **21**, 191–193
Kingsbury, D. T. (1969) J. Bacteriol. **98**, 1400–1410
Kingsman, A. J., Gimlich, R. L., Clarke, L., Chinault, A. C. & Carbon, J. (1981) J. Mol. Biol. **145**, 619–632
Kleckner, N. (1977) Cell **11**, 11–23
Kleckner, N. (1981) Annu. Rev. Genet. **15**, 341–404
Koch, A. L. (1972) Genetics **80**, 279–316
Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E. & Cech, T. R. (1982) Cell **31**, 147–157
Kugimiya, W., Ikenaga, H. & Saigo, K. (1983) Proc. Natl. Acad. Sci. U.S.A. **80**, 3193–3194
Laird, C. D. (1971) Chromosoma **32**, 378–406
Lauth, M. R., Spear, B. B., Heumann, J. & Prescott, D. M. (1976) Cell **7**, 67–74
Leach, D. R. F. & Stahl, F. W. (1983) Nature (London) **305**, 448–451
Lerman, M. I., Thayer, R. E. & Singer, M. F. (1983) Proc. Natl. Acad. Sci. U.S.A. **80**, 3966–3970
Levis, R. & Rubin, G. M. (1982) Cell **30**, 543–550
Liao, L. W., Rozenzweig, B. & Hirsh, D. (1983) Proc. Natl. Acad. Sci. U.S.A. **80**, 3585–3589
Lonberg, N. & Gilbert, W. (1985) Cell **40**, 81–90
Lueders, K. K. & Kuff, E. L. (1977) Cell **12**, 963–972
McGinnis, W., Shemoen, A. W. & Beckendorf, S. K. (1983) Cell **34**, 75–84
McLintock, B. (1956) Cold Spring Harbor, Symp. Quant. Biol. **21**, 197–216
Macreadie, I. G., Scott, R. M., Zinn, A. R. & Butow, R. A. (1985) Cell **41**, 395–402
Majors, J. E. & Varmus, H. E. (1981) Nature (London) **289**, 253–258
Manning, J. E., Schmid, C. W. & Davidson, N. (1975) Cell **4**, 141–155
Martin, S. L., Voliva, C. F., Burton, F. H., Edgell, M. H. & Hutchinson, C. A., III (1984) Proc. Natl. Acad. Sci. U.S.A. **81**, 2308–2312
Mazrimas, J. A. & Hatch, F. T. (1972) Nature (London) New Biol. **240**, 102–105
Mellor, J., Fulton, S. M., Dobson, M. J., Wilson, W., Kingsman, S. M. & Kingsman, A. J. (1985) Nature (London) **313**, 243–246
Modolell, J., Bender, W. & Meselson, M. (1983) Proc. Natl. Acad. Sci. U.S.A. **80**, 1678–1672
Moyzis, R. K., Bonnet, J., Li, D. W. & T'so, P. O. P. (1981*a*) J. Mol. Biol. **153**, 841–870
Moyzis, R. K., Bonnet, J., Li, D. W. & T'so, P. O. P. (1981*b*) J. Mol. Biol. **153**, 871–876
Nader, W. F., Edlind, T. D., Huettermann, A. & Sauer, H. W. (1985) Proc. Natl. Acad. Sci. U.S.A. **82**, 2698–2702
Nevers, P. & Saedler, H. (1977) Nature (London) **268**, 109–115
Nunberg, J. H., Kaufman, R. J., Chang, A. C. Y., Cohen, S. N. & Schimke, R. T. (1980) Cell **19**, 355–364
O'Hare, K. & Rubin, G. M. (1983) Cell **34**, 25–35
Olmo, E. & Morescalchi, A. (1975) Experientia **31**, 804–806
Orgel, L. E. & Crick, F. H. C. (1980) Nature (London) **284**, 604–607
Owens, G. P., Chaudhari, N. & Hahn, W. E. (1985) Science **229**, 1263–1265
Palm, D., Goerl, R. & Burger, K. J. (1985) Nature (London) **313**, 500–502
Pearston, D. H., Gordon, M. & Hardman, N. (1985) EMBO J. **4**, 3557–3562
Perlman, J. (1983) Gene **21**, 87–94
Perlman, S., Phillips, C. A. & Bishop, J. D. (1976) Cell **8**, 33–42
Peoples, O. P. & Hardman, N. (1983) Nucleic Acids Res. **11**, 7777–7788
Peoples, O. P., Robinson, A. C., Whittaker, P. A. & Hardman, N. (1983) Biochim. Biophys. Acta **741**, 204–213
Peoples, O. P., Whittaker, P. A., Pearston, D. H. & Hardman, N. (1985) J. Gen. Microbiol. **131**, 1157–1165
Potter, S. S., Truett, M., Phillips, M. & Maher, A. (1980) Cell **20**, 639–647
Reeder, R. H. & Roan, J. G. (1984) Cell **38**, 39–44
Rees, H. & Jones, R. N. (1967) Nature (London) **216**, 825–826
Rogers, J. H. (1984) Int. Rev. Cytol. **93**, 187–279
Rogers, J. H. (1985) Biochim. Biophys. Acta **824**, 113–120
Rothfels, K., Sexsmith, E., Heimburger, M. & Krause, M. O. (1966) Chromosoma **20**, 54–74
Rubin, G. M., Brorein, W. J., Jr., Dunsmuir, P., Flavell, A. J., Levis, K., Strobel, E., Toole, J. J. & Young, E. (1981) Cold Spring Harbor Symp. Quant. Biol. **45**, 619–628
Rubin, G. M., Kidwell, M. G. & Bingham, P. M. (1982) Cell **30**, 987–994
Saigo, K., Kugimiya, W., Matsuo, Y., Inouye, S., Yoshioka, K. & Yuki, S. (1984) Nature (London) **312**, 659–661
Scherer, G., Tschudi, C., Perera, J., Delius, H. & Pirrotta, V. (1982) J. Mol. Biol. **157**, 435–451
Schindler, C. W. & Rush, M. G. (1985) J. Mol. Biol. **181**, 161–173
Schmid, C. W. & Deininger, P. L. (1975) Cell **6**, 345–358
Schmid, C. W., Manning, J. E. & Davidson, N. (1975) Cell **5**, 159–172
Sharpe, P. A. (1983) Nature (London) **301**, 471–472
Shafit-Zagardo, B., Brown, F., Zarodny, P. & Maio, J. (1983) Nature (London) **304**, 277–280
Shepherd, N. S., Schwarz-Sommer, Z., Blumberg vel Spalve, J., Gupta, M., Wienand, U. & Saedler, H. (1984) Nature (London) **307**, 185–187
Sherratt, D. J., Arthur, A. & Burke, M. (1981) Cold Spring Harbor Symp. Quant. Biol. **45**, 275–282
Shiba, T. & Saigo, K. (1983) Nature (London) **302**, 119–124
Shoemaker, C., Hoffman, J., Goff, S. P. & Baltimore, D. (1981) J. Virol. **40**, 164–172
Sinclair, J. H., Sang, J. H., Burke, J. F. & Ish-Horowicz, D. (1983) Nature (London) **306**, 198–200
Singer, M. F. (1982) Cell **28**, 433–434
Singer, M. F. & Skowronski, J. (1985) Trends Biochem. Sci. **10**, 119–122
Skowronski, J. & Singer, M. F. (1985) Proc. Natl. Acad. Sci. U.S.A., **82**, 6050–6054
Smith, G. P. (1973) Cold Spring Harbor Symp. Quant. Biol. **38**, 507–514
Smith, G. P. (1976) Science **191**, 528–535
Smith, G. P. (1978) Trends Biochem. Sci. **3**, N34–N36
Smith, G. R., Kunes, S. M., Schultz, D. W., Taylor, A. & Triman, K. L. (1981) Cell **24**, 429–436
Sobieski, D. A. & Eden, F. C. (1981) Nucleic Acids Res. **9**, 6001–6015
Southern, E. M. (1975) J. Mol. Biol. **94**, 51–69
Spradling, A. C. & Rubin, G. M. (1981) Annu. Rev. Genetics **15**, 219–264
Stansfield, S. W. & Lengyel, J. A. (1979) Proc. Natl. Acad. Sci. U.S.A. **76**, 6142–6146
Starlinger, P. (1984) Trends Biochem. Sci. **9**, 125–127
Stone, E. M., Rothblum, K. N. & Schwartz, R. J. (1985) Nature (London) **313**, 498–500
Strauss, N. A. (1971) Proc. Natl. Acad. Sci. U.S.A. **68**, 799–802
Summers, J. & Mason, W. S. (1982) Cell **29**, 403–415
Sutcliffe, J. G., Milner, R. J., Gottesfeld, J. M. & Lerner, R. A. (1984) Nature (London) **308**, 237–241
Temin, H. M. (1980) Cell **21**, 599–600
Temin, H. M. (1981) Cell **27**, 1–3
Temin, H. M. (1982) Cell **28**, 3–5
Thomas, C. A., Jr. (1971) Annu. Rev. Genet. **5**, 237–256
Timberlake, W. E. (1978) Science **202**, 973–974
Toh, H., Kikuno, R., Hayashida, H., Miyata, T., Kugimiya, W., Inouye, S., Yuki, S. & Saigo, K. (1985) EMBO J. **4**, 1267–1272
Ullu, E. & Tschudi, C. (1984) Nature (London) **312**, 171–172
Van Arsdell, S. W., Denison, R. A., Bernstein, L. B. & Werner, A. M. (1981) Cell **26**, 11–17
Varmus, H. E. (1982) Science **216**, 812–820
Voliva, C. F., Jahn, C. L., Comer, M. B., Hutchison, C. A., III & Edgell, M. H. (1983) Nucleic Acids Res. **11**, 8847–8859
Von der Helm (1977) Proc. Natl. Acad. Sci. U.S.A. **74**, 911–915

Waalwyjk, C. & Flavell, R. A. (1978) Nucleic Acids Res. **5**, 4631–4640
Walter, P. & Blobel, G. (1980) Proc. Natl. Acad. Sci. U.S.A. **77**, 7112–7116
Weiss, R., Teich, N., Varmus, H. E. & Coffin, J. (eds.) (1982) Molecular Biology of Tumour Viruses: RNA Tumour Viruses, Cold Spring Harbor Laboratory, New York
Wensink, P. C., Finnegan, D. J., Donelson, J. E. & Hogness, D. S. (1974) Cell **3**, 315–325
Wensink, P. C., Tabata, S. & Pachl, C. (1979) Cell **18**, 1231–1248
Wilson, D. A. & Thomas, C. A., Jr. (1974) J. Mol. Biol. **84**, 115–144
Wyman, A. R., Wolfe, L. B. & Botstein, D. (1985) Proc. Natl. Acad. Sci. U.S.A. **82**, 2880–2884
Young, M. W. (1979) Proc. Natl. Acad. Sci. U.S.A. **76**, 6274–6278
Zachar, Z. & Bingham, P. M. (1982) Cell **30**, 529–541