Hum. Genet. 55, 1–18 (1980)

Human Genetics

Review Articles

Sequence Organization of Animal Nuclear DNA

J. Schmidtke and J. T. Epplen*

Institut für Humangenetik der Universität, Nikolausberger Weg 5a, D-3400 Göttingen, Federal Republic of Germany

Summary. Animal nuclear genomes contain DNA sequences of various degrees of repetition. These sequences are organized in highly ordered fashions; repetitive and nonrepetitive sequences either alternate in short periods, i.e., short [0.2—0.4 kilobases (kb) long] repeats are flanked by nonrepetitive sequences less than 2 kb long, or in longer periods, with repetitive and/or nonrepetitive sequences extending for several kilobases. There are two main categories of genome organization, namely those exhibiting short-period interspersion and those that do not. There are arguments for and against a regulatory role of short interspersed repetitive sequences.

Besides the merely 'statistical' kinetic approach by conventional reassociation kinetics, sequence organization has been studied by restriction endonuclease mapping and nucleotide sequencing. Such studies have revealed some general features of the organization of the eukaryotic gene and its transcripts, namely possible 'promoters', 'leaders', 'introns', 'exons', 'flanking sequences', 'caps', ribosome-binding sites, and poly(A) sequences. This paper discusses how these elements of a gene might serve regulatory roles in its expression.

Introduction

One of the most prominent features of eukaryotic DNA is the presence of repeated sequences (Britten and Kohne, 1968). These sequences appear to be distributed in nonrandom fashion within the genome, and show an extraordinarily similar length distribution and arrangement in most eukaryotes, including such diverse organisms as dinoflagellates and man. These findings have supported an early speculation about a functional role of these elements in gene expression in higher genomes. Recently, however, a fairly large amount of information on alternative DNA sequence organization patterns has accumulated, characterized by a strikingly different distribution of repetitive sequence length and by a basically different mode of interspersion with nonrepetitive DNA stretches. These findings have raised many questions about the function and evolution of the genomes of higher organisms. It is the purpose of this article to review the findings on contrasting sequence organization patterns in animal genomes and to discuss the implications of such findings, together with recent knowledge about gene structure, for the suspected role of certain DNA sequences in gene regulation.

Sequence Composition of Animal Genomes

Repetitive elements typically constitute 30% of animal genomes investigated over a wide range of size (Fig. 1). This means that single-copy DNA always appears to account for a constant proportion of the genome, regardless of the biological complexity of the organism concerned (Davidson et al., 1975a). A similar situation is encountered in plant genomes, but here repetitive sequences make up a much larger proportion of genomic DNA, usually about 70% (Thompson, 1978). It has been suggested that increases in total genome size reflect periodic additions of repeated sequences, at rates different in animals and plants, while single-copy DNA accumulates by gradual evolutionary divergence of these repeats. According to this view, a large part of repeated as well as nonrepeated DNA does not serve any sequence-dependent function ['secondary DNA' in the terminology of Hinegardner (1976); 'nucleoskeletal'

Offprint requests to: J. Schmidtke

* *Present address:* City of Hope Medical Center, Division of Biology, Duarte, CA, USA

0340-6717/80/0055/0001/$ 03.60

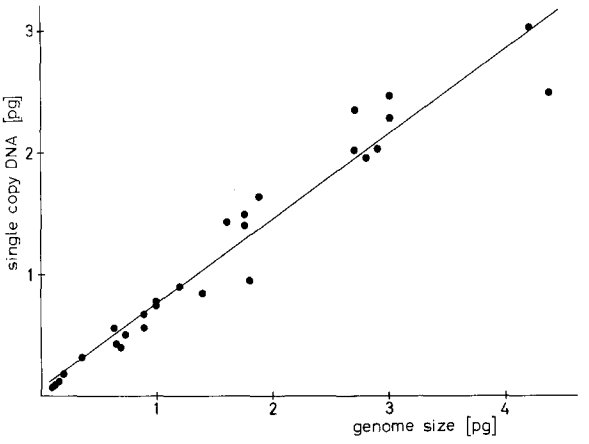

Fig. 1. Amount of single-copy DNA in animal genomes. Values for single-copy DNA content are plotted as a function of genome size. Linear regression analysis ($r = 0.9732$) shows that single-copy DNA constitutes 70% of the genome on average, irrespective of the absolute size of the genome. Only single-copy DNA measurements were used, which are corrected for the effect of sequence interspersion. Data are from Angerer et al. (1975), Beauchamp et al. (1979), Christie and Skinner (1979), Crain et al. (1976), Davidson et al. (1973), Eden et al. (1978), Efstratiadis et al. (1976), Epplen et al. (1978, 1979), Ginelli et al. (1977), Goldberg et al. (1975), Graham et al. (1974), Manning et al. (1975), Moritz and Roth (1976), Pearson et al. (1978), Schmid and Deininger (1975), Schmidtke et al. (1979a), Smith and Boal (1978), Wells et al. (1976), and this paper

DNA in that of Cavalier-Smith (1978), who assumes that a large part of the genomic DNA has a functional role in nuclear statics, perhaps also by coding for a 'nucleoskeletal' RNA]. In consequence, genomic rearrangement leading to extensive interspersion of repeated and nonrepeated sequences may similarly be the result of stochastic processes and not of selection for genotypic or phenotypic functions (Thompson, 1978).

Although repeated sequences often display a wide spectrum of repetition frequency in a genome, at least for descriptive purposes, they are customarily conceived as discrete genome components. If genomic DNA is analyzed by denaturation–renaturation techniques, at least three theoretical kinetic components are normally distinguishable: a very rapidly renaturing one, a fast renaturing one, and a slowly renaturing component.

The very rapidly renaturing DNA component frequently consists of sequences repeated an enormous number of times, in the order of 10^6 per genome. These sequences may often be found in the satellite DNA bands observed in equilibrium density gradient centrifugation. On the chromosomes these sequences often occur in a clustered arrangement (recent studies of human satellite DNA: Miklos and John, 1979; Cooke and Hindley, 1979). Possible functions of satellite DNA have recently been discussed (Singh et al., 1976; Jones, 1978; Vogel and Motulsky, 1979). In addition, the very rapidly renaturing fraction may also contain foldback (palindromic) sequences, i.e., sequences capable of forming intrastrand duplexes with very fast (first-order) kinetics. In human and *Drosophila* DNA, for example, these structures form 6% of the genomes (Schmid et al., 1975; Dott et al., 1976). Self-complementary sequences are often found in the 5′ region of eukaryotic and prokaryotic genes (see Nishioka and Leder, 1979). Short palindromes serve as recognition sites for many bacterial restriction endonucleases (Smith, 1979); this fact has greatly facilitated the analysis of genome structure on the sequence level (see below).

The DNA renaturing with intermediate rate contains sequences repeated in the order of 10^1—10^5 times. This class may sometimes be clearly subdivided into distinct kinetic subclasses, such as in the case of the human genome (Schmid and Deininger, 1975), the mollusc *Aplysia* (Angerer et al., 1975), and a number of teleosts (Schmidtke et al., 1979c). It is sometimes observed that among closely related species with varying genomic DNA content, this kinetic class accounts for a large part of the genome size differences (Schmidtke et al., 1979b). In the typical case, sequences of this class are extensively interspersed with nonrepetitive elements.

The slowly renaturing genome component contains sequences that are usually present only once per haploid genome (single-copy, 'unique' DNA). Since 'classic genes' are usually present only once per genome, this component contains most of the structural gene sequences. These sequences however, constitute only a very small proportion of the total unique DNA. Various factors suggest that the number of genes is in the order of a few times 10^4 (Ohno, 1971; Kiper, 1979), which means that in the human genome, for example, less than 5% of the DNA present in single-copy consist of classic genes. It should be pointed out, however, that a sequence could be scored as single-copy by the kinetic approach although it is in fact repetitive. This is the case when renatured repetitive duplexes are grossly mismatched due to extensive base sequence divergence, because such mispairing reduces the reassociation rate (Bonner et al., 1973). The existence of such a class of low-repetition sequences, renaturing with single-copy properties in the human genome, has recently been described (Deininger and Schmid, 1979). Nevertheless, a large proportion of the genome is present in single copy and does not serve any known sequence-dependent function.

In general, all cell nuclei in a multicellular organism are thought to contain identical genomes with regard to sequence composition and arrangement. Several notable exceptions exist: for example, at the sites of

chromosomal puffs in *Rhynchosciara* extra DNA synthesis occurs (Breuer and Pavan, 1955); in somatic cell nuclei of *Ascaris,* certain genome portions are lost (Boveri, 1887; Roth, 1979); amplification of specific DNA sequences has been observed in the differentiating chicken cartilage and neural retina (Strom et al., 1978); and DNA sequences appear to be translocated during embryogenesis of the sea urchin (Dickinson and Baker, 1978) and during the differentiation of mouse lymphocytes (Brack et al., 1978). The relevance of these findings for cellular differentiation is not always clear; it is doubtful whether changes of DNA sequence composition and organization are mechanisms generally employed in the differentiating cell.

Patterns of Sequence Arrangement in Animal Genomes

The genomic arrangement of repetitive and nonrepetitive sequences, as studied by renaturation techniques, has been described in terms of approximate sequence lengths. Since these sequences are *kinetically* defined, sequences terminate when they border sequences of a different degree of repetition. This could mean that a single nonrepetitive sequence, for example, is *functionally* subdivided (e.g., exon–intron, see below); likewise, repetitive sequences may occur as 'isolated' or as 'clustered' repeats. Alternation of repetitive and nonrepetitive sequences has also been termed 'interspersion' (Davidson et al., 1973). Interspersion of short (0.2—0.4 kb) repetitive sequences with short (< 2 kb) nonrepetitive sequences has been found to be a prominent feature of most eukaryotic genomes (Fig. 2, type A). In addition, interspersion of longer (several kilobases) nonrepetitive sequences with repetitive sequences of undefined length occurs in smaller parts of most genomes (Fig. 2, type B). Finally, a 'long-period' sequence arrangement exists, in which both the nonrepetitive and the repetitive sequences extend uninterrupted for several kilobases (Fig. 2, type C).

a) Genomes with Short-period DNA Sequence Interspersion

Genomes that contain the type A sequence arrangement of Fig. 2 are said to be organized in a 'short-period DNA sequence interspersion,' although they may contain the other types of sequence arrangement as well. This is because most information on sequence organization is derived from experiments comparing the renaturation kinetics of DNA of different fragment length. One such experiment is depicted in Fig. 3. The reassociation of DNA molecules a few thousand nucleotides long appears considerably accelerated compared with molecules a few hundred nucleotides in length. In the hydroxyapatite (HAP) assay used in this experiment, a molecule is scored as 'reassociated' even if it is only partly in duplex. The increase of DNA binding to HAP is mostly due to the fact that unreassociated single-stranded stretches are trapped together with reassociated neighboring regions. If nonrepetitive DNA is largely interspersed with repetitive DNA, as in the type A arrangement of Fig. 2, the curves of Fig. 3 depict the expected result of the reassociation experiment. This type of sequence organization is often also referred to as the '*Xenopus*-pattern,' from the organism in which it was first studied in detail (Davidson et al., 1973). However, there is some ambiguity about this term, because it is also used to describe genomes in which types A and B of sequence arrangement coexist.

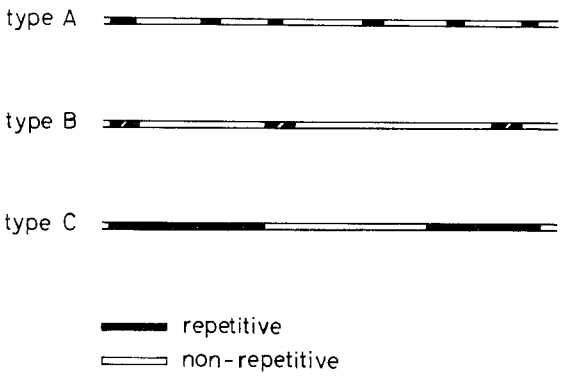

Fig. 2. Schematic presentation of the types of interspersion of repetitive with nonrepetitive sequences observed in eukaryotic genomes. Repetitive sequences are indicated as *full* stretches, nonrepetitive sequences as *empty* stretches. In the type A arrangement repetitive sequences are about 0.3 kb long; their length is not defined in the type B arrangement; and they extend for several kilobases in the type C arrangement. Nonrepetitive sequences extend for less then 2 kb in the type A sequence arrangement, and for several kb in types B and C

In a typical Xenopus-like organized genome the type A arrangement accounts for about half of the genomic DNA. For example, about 52% of the human genome consists of an interspersion of single-copy sequences averaging 2 kb, with repetitive sequences of about 0.4 kb in length (Schmid and Deininger, 1975). In the sea urchin and in *Xenopus,* about 55% of the genome consist of single-copy stretches 0.9 kb long spacing 0.4 kb repetitive elements (Davidson et al., 1973; Graham et al., 1974). In genomes characterized by short-period interspersion, a significant amount of the DNA is arranged in the type B sequence arrangement (Fig. 2). In the *Xenopus* and sea urchin genomes this portion makes up about 20% of the DNA or about 30% of the nonrepetitive sequence, which occurs in stretches longer than 3.4 kb (Davidson et al., 1973; Graham et al., 1974). In the rat genome, the long inter-

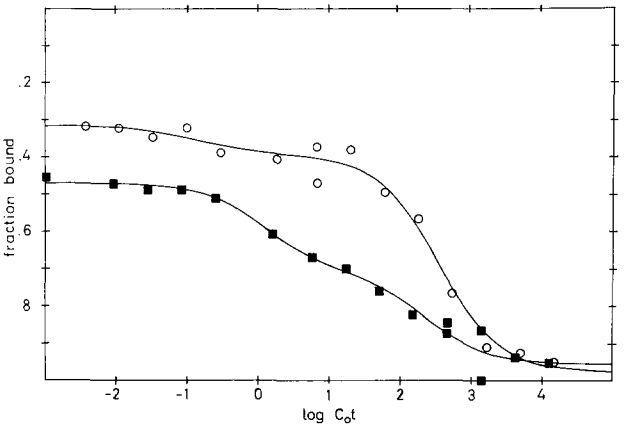

Fig. 3. An example of short-period DNA sequence arrangement: genome of the herring *(Clupea harengus)*. DNA was sheared by sonication to 240 (○) and 1800 (■) nucleotides, dissociated and reassociated in 0.04–0.55 *M* phosphate buffer at 60–65°C. The fraction of DNA fragments containing reannealed regions after renaturation to the desired value of C_0t (expressed equivalent to 0.12 *M* phosphate buffer (Britten et al., 1974) was assayed by hydroxyapatite chromatography. The curves represent computer-generated best fits to the data points for three kinetic components of unrestricted rate constants. In the 240 nucleotide curve these components are: 32% highly repetitive (plus fold-back) with a rate constant $K > 10^3$ (moles of nucleotides/liter)$^{-1} \cdot s^{-1}$, 8% intermediate repetitive with a $K = 6.9\,M^{-1} \cdot s^{-1}$, and 58% slow (single-copy) with a $K = 0.0028\,M^{-1} \cdot s^{-1}$. In the 1800 nucleotide reassociation curve both the highly repetitive and the intermediately repetitive kinetic components appear increased, to 47% and 24%, respectively, whereas the slow component is decreased to 25%. This shows that on the fragment length used, kinetically distinct sequences are highly interspersed with each other

spersion period appears to be in excess of 6 kb and to comprise 10%—15% of the DNA (Pearson et al., 1978). Other animal genomes, which have been investigated in detail, e.g., *Antheraea* (Efstratiadis et al., 1976) and man (Schmid and Deininger, 1975) show a comparable amount and mode of spacing of long period sequence arrangement.

A great deal of attention has been given to the length distribution of repetitive sequences in genomes predominantly organized in short periods. Besides the short 0.2—0.4 kb repeats, longer repetitive elements occur in every genome investigated. These elements are generally longer than 1.5 kb (the exclusion limit of the commonly employed agarose A-50 chromatography) and account for almost 50% of the repeated DNA on average, but may comprise as little as 25%, as in *Antheraea* (Efstratiadis et al., 1976), or even as much as 65%, as in *Crassostrea* (Goldberg et al., 1975). There is, however, no simple correlation of the fraction of the repetitive DNA in long arrays and the genome size (Davidson et al., 1975). The amount of short repetitive segments even in genomes with an exceptionally low content of these elements, such as *Pisaster ochraceous* (Smith and Boal, 1978), is always sufficient to account for at least the major fraction of single-copy DNA present in a short interspersion pattern.

Reports as to the extent of sequence homology between long and short repeat elements within one and the same genome are conflicting: In the rat, long and short sequences partially overlap (Wu et al., 1977); in sea urchin DNA these elements are essentially distinct, according to Eden et al. (1977), or share common stretches, according to Chaudhari and Craig (1979).

It has been suggested that some of the long repeats generally represent clusters of shorter subrepeats (Rimpau et al., 1978; Chaudhari and Craig, 1979), or satellite DNA sequences (Britten et al., 1976). In the human genome, a major portion of the long repetitive sequences has been shown to consist of satellite sequences (Houck et al., 1978). Also occurring in a clustered arrangement (e.g., in *Xenopus*) are the repetitive genes for high-molecular-weight ribosomal RNA, 5.8S and 5S ribosomal RNA, transfer RNA, histones (for review see Bostock and Sumner, 1978), and probably other transcribed multigene families (Galau et al., 1976a). The recent findings of Wensink et al. (1979) on the subrepeat organization of long repetitive sequences in *Drosophila* DNA are discussed below.

Fig. 4 illustrates that the short period interspersion pattern is widespread among animal genomes. It occurs in both the protostome and deuterostome branches of the animal kingdom. It should be noted that short-period DNA sequence interspersion is also found in protists, such as *Euglena gracilis,* a unicellular alga (Rawson et al., 1979), *Crypthecodinium cohnii,* a dinoflagellate (Allen et al., 1975), and the ciliate *Tetrahymena pyriformis* (Borchsenius et al., 1978). In the last organism, however, only 15% of the genome is organized in short-period interspersion, whereas 80% of the unique sequences extend in long stretches uninterrupted by repeats. Short-period interspersion is also typical for plants [for recent reviews see Thompson (1978) and Herzfeld and Kiper (1979)], including the primitive fungus *Dictyostelium discoideum* (Firtel and Kindle, 1975). In addition, in most plant genomes showing short-period interspersion, substantial amounts of DNA occur in long, single-copy stretches (Herzfeld and Kiper, 1979).

A peculiar form of short period sequence arrangement has been observed in the crab *Geryon quinquedens* (Christie and Skinner, 1979). While in all other genomes showing short-period interspersion the repeat elements interspersed with single-copy DNA are intermediately repetitive, in the red crab unique stretches are contiguous with highly repeated sequences.

Another different form of genome organization still has been found in the case of the albatross *(Diomedea*

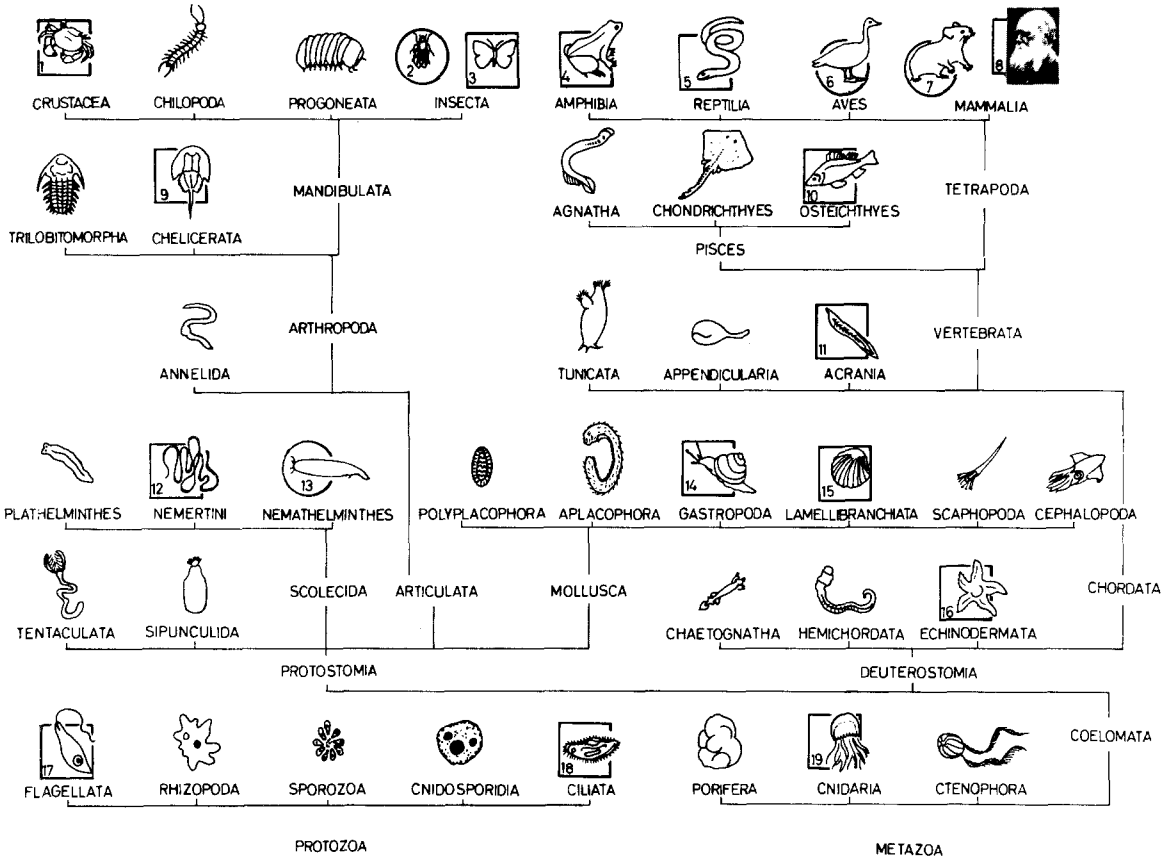

Fig. 4. Occurrence of contrasting DNA-sequence interspersion patterns in animals and protozoans. The classification scheme followed Remane et al. (1976). Species representing a taxonomic group in which short-period sequence arrangement occurs are boxed, whereas those representing a group characterized by long-period arrangement are encircled. It is obvious that short-period interspersion is prevalent, but long-period interspersion is the exclusive organization type in several diverse lineages. Species names and references: 1, *Geryon quinquedens* (Christie and Skinner, 1979); 2, *Drosophila melanogaster* (Manning et al., 1975; Crain et al., 1976b); *Chironomus tentans* (Wells et al., 1976), *Apis mellifera* (Crain et al., 1976a); *Sarcophaga bullata* (Samols and Swift, 1979); 3, *Musca domestica* (Crain et al., 1976a); *Antheraea pernyi* (Efstratiadis et al., 1976); 4, *Xenopus laevis* (Davidson, 1973; Davidson et al., 1973; Chamberlin et al., 1975; Baldari and Amaldi, 1976, 1977); *Bufo bufo, Triturus cristatus, Necturus maculosus* (Baldari and Amaldi, 1976, 1977); 5, *Python reticularis, Caiman crocodilus, Terrapene carolina triungius* (Epplen et al., 1979); 6, *Gallus domesticus* (Arthur and Straus, 1978; Eden and Hendrick, 1978; Epplen et al., 1978); *Cairina moschata domestica* (Epplen et al., 1978); *Columba livia domestica* (Epplen et al., 1979); 7, *Mesocricetus auratus* (Moyzis et al., 1977); 8, *Homo sapiens* (represented by Charles Darwin) (Schmid and Deininger, 1975; Deininger and Schmid, 1976; Ginelli and Corneo, 1976); *Bos taurus* (Britten and Smith, 1970); *Mus musculus* (Ginelli et al., 1977); *Rattus norvegicus* (Pearson et al., 1978; Wilkes et al., 1978); 9, *Limulus polyphemus* (Goldberg et al., 1975); 10, *Clupea harengus* (this paper); 11, *Branchiostoma lanceolatum* (Schmidtke et al., 1979a); 12, *Cerebratulus lacteus* (Goldberg et al., 1975); 13, *Panagrellus silusiae* (Beauchamp et al., 1979); *Ascaris suum* (Roth, 1979); 14, *Aplysia californica* (Angerer et al., 1975); 15, *Crassostrea virginica, Spisula solidissima* (Goldberg et al., 1975); 16, *Strongylocentrotus purpuratus* (Graham et al., 1974; Lee et al., 1977); *Pisaster ochraceus* (Smith and Boal, 1978); *Strongylocentrotus intermedius, S. nudus, Scaphechinus mirabilis, Echinocardium cordatum, Asterias amurensis, Distelasterias nipon, Evasterias retifera, Stichopus japonicus* (Brykov et al., 1978); 17, *Crypthecodinium cohnii* (Allen et al., 1975); *Euglena gracilis* (Borchsenius et al., 1978); *Aurelia aurita* (Goldberg et al., 1975)

epomophora) and the guillemot *(Uria aalge)* (Ginatulin and Ginatulina, 1979, and personal communication). In these bird genomes a class of short (~0.3 kb) repeat sequences exists, but these are linked to still shorter (~0.15 kb) single-copy elements.

b) *Genomes Without Short-period DNA Sequence Interspersion*

For some time it was generally believed that only a few insect genomes apparently lack short period DNA sequence interspersion, namely *Drosophila melanogaster* (Manning et al., 1975; Crain et al., 1976a), *Chironomus tentans* (Wells et al., 1976), and *Apis mellifera* (Crain et al., 1976b). Soon afterwards, however, it was recognized that the apparent lack of short-period interspersion is much more widespread. One mammal (Moyzis et al., 1977), several bird genomes (Arthur and Straus, 1978; Eden and Hendrick, 1978; Epplen et al., 1978, 1979), a nematode (Beauchamp et al., 1979), several fungi (Hudspeth et al., 1977; Timberlake, 1978; Har-

shey, 1979), and one higher plant (Wimpee and Rawson, 1979) are typified by the alternation of both single-copy and repetitive sequences in longer stretches (type C sequence arrangement of Fig. 2). In these genomes no discrete class of 0.2—0.4 kb long interspersed intermediate repeat elements could be demonstrated by renaturation experiments. In *Drosophila* the average length of interspersed middle repetitive sequences is 5.6 kb, and the length of the nonrepetitive region linking the repeats is greater than 13 kb; at least one-third, possibly all, of the middle repetitive sequences are organized in this way, but it is also possible that some of the intermediate repeats occur in even larger stretches (Manning et al., 1975). The results of Crain et al. (1976a) substantially agree with these findings.

The organization of the genome of the closely related species *Chironomus tentans* is very similar. In this species, in at least 35% of the genome, 4.3 kb-long repeats alternate with 13.5 kb-long nonrepetitive sequences; alternatively these repeats are distributed throughout the genome with an inter-repeat distance of 36 kb (Wells et al., 1976). In the Syrian hamster the length of interspersed intermediate repeats is also 4—5 kb, but the interrupting single-copy sequences are only 7 kb long (Moyzis et al., 1977). In the chicken genome, repeat elements average 2—4 kb, while single-copy spacers are about 4.5 kb in length in 40% of the genome and even longer ones occur in the remaining part of the genome (Arthur and Straus, 1978; Eden and Hendrick, 1978). In the nematode *Panagrellus silusiae,* repeated sequences are estimated to be as much as 10 kb long (Beauchamp et al., 1979). Sequence length estimates cannot be given for the remaining animal genomes, because these have not yet been studied in sufficient detail. It can be seen from these studies that the sequence lengths vary considerably among different genomes; it is emphasized, however, that a feature common to all genomes characterized by long-period interspersion is the apparent absence of 0.2—0.4 kb long interspersed repeated sequences. The possibility cannot be excluded that these elements are so rare in these genomes that they cannot be detected by reassociation methods. It appears, however, that the vast majority of short repeated sequences are dispensible and play no obvious general role in eukaryote genome function.

Wensink et al. (1979) have recently shown that at least some of the apparent long stretches of moderately repeated DNA in the *Drosophila* nuclear genome actually represent large clusters of densely spaced, short (< 1 kb) repetitive subsequences. This study also revealed that different clusters have some elements in common, but arranged in different orders. The authors speculate that the subrepeat sequences are transposable elements.

It can be seen from Fig. 4 that the long-period interspersion pattern, although much less frequently observed than the short-period pattern, is no less widespread than the latter in the animal kingdom. While there is no apparent correlation between phylogenetic position and the type of DNA sequence arrangement, the latter is clearly correlated with genome size in that smaller genomes tend to be arranged in long periods. This is most impressively documented in cases where related species show contrasting DNA sequence organization. For example, *Drosophila melanogaster, Chironomus tentans, Apis mellifera* and *Sarcophaga bullata,* all of which are insects lacking short-period interspersion, have genome sizes of 0.1—0.61 pg, while the two insects *Musca domestica* and *Antheraea pernyi* (which are endowed with short-period interspersion) have genome sizes of 0.9 and 1.0 pg, respectively. The two nematodes *Panagrellus silusiae* and *Ascaris suum* (long period interspersion) have genomes as small as 0.097 and 0.25 pg, respectively, while the nemertean worm *Cerebratulus lacteus* (short-period interspersion) has a genome 5—14 times larger. Fungi, which have very small genomes in comparison with other eukaryotes, tend to be organized in long periods, and the pearl millet, a higher plant with an unusually low DNA content, also lacks short-period interspersion. Birds as a group have smaller genomes than their closest relatives, the reptiles, and in fact most bird genomes are organized in long period and all reptiles investigated so far in short period. It has been noted, however, that among birds an inverse relationship exists between genome size and the amount of single-copy DNA interspersed with repeat elements on a given DNA fragment length (Epplen et al., 1979). However, these findings are preliminary and await further confirmation. The only marked exception from the apparent rule that organisms at the lower end of the genome size scale of a related group of species tend to lack short-period interspersion is the mammals. *Mesocricetus auratus* has been reported to have a genome organized in the long-period fashion, but it has a genome twice as large as that of mouse (Bachmann, 1972), whose genome is arranged in short periods (see: Note added in proof).

The general correlation between small genome size and lack of short-period interspersion is consistent with the view of DNA loss in an evolutionary transition from the short to the long interspersion pattern (Crain et al., 1976b). Species lacking short-period interspersion have been seen to contain a relatively small amount of repetitive DNA, but a possible loss of repeated sequences during evolution of these genomes could only account for the formation of long single-copy arrays. Generation of the long repeat sequences from shorter ones can only be explained by a conco-

mitant loss of single-copy spacers (Davidson et al., 1977). It should be noted that the recent findings of Wensink et al. (1979) on the subrepeat structure of the long repeats is compatible with this view. Loss of genomic DNA along with specialization during evolution has long been noted (for review see Hinegardner, 1976); Britten and Davidson (1971) and Davidson et al. (1977) have argued that both single-copy sequences and interspersed repeat sequences would serve as elements promoting genomic rearrangement, including rearrangement of control regions, a prerequisite for evolutionary change (e.g., King and Wilson, 1975). Therefore, genomes typified by long-period sequence arrangement may derive from genomes organized in short period in the course of specialization; genomes that maintain a short-period interspersion pattern, on the other hand, are the ones that retain evolutionary flexibility.

Apart from their possible role in genome reorganization events, it has been suggested that the interspersed repeated sequences play an important part in the regulation of DNA replication (Mattern and Painter, 1977; Cavalier-Smith, 1978), and most notably in gene expression (Britten and Davidson, 1969, 1971; Davidson et al., 1977; Davidson and Britten, 1979). The next section summarizes this last hypothesis and the experimental evidence purporting to support it, together with findings apparently in conflict with this view.

A Suspected Functional Role of Short Interspersed Repeats

A fundamental difference in the basic design of prokaryote and eukaryote genome organization is the physical localization of structural genes that are functionally related. In the bacterial genome, genes coding for proteins involved in a common synthetic pathway are usually linked physically ('operon'), and the primary transcript is read off as a giant polycistronic messenger RNA. In the eukaryote, on the other hand, genes that are functionally related are usually physically dispersed throughout the genome. In the human genome, for example, genes coding for enzymes catalyzing the first six main steps of glucose metabolism map on six different chromosomes (see Donald and Hamerton, 1978). Genes that control steps of cell differentiation in a multicellular organism are similarly expected to be generally nonlinked. It is apparent that a model that is suitable for description of transcription level control of gene expression must show how genes that are located on different genomic sites are regulated together. Britten and Davidson (1969, 1971) and Davidson and Britten (1973, 1979) have proposed a model to account for the coordinate regulation of the noncontiguous structural gene sequences in higher organisms, based on control events occurring both transcriptionally and posttranscriptionally.

Relying on the concept of Jacob and Monod (1961), according to which recognition elements map adjacent to structural genes they control, Britten and Davidson originally (1969) suggested that gene batteries that become activated together share homologous repetitive 'receptor' sequences. These elements would serve as binding sites for diffusible sequence-specific 'activator' (RNA) molecules. Activator RNA is transcribed from 'integrator' genes, which in turn are controlled by 'sensor' sequences, to which agents that induce the occurrence of specific patterns of genome activity, such as hormones would 'bind.' In keeping with the logic of this model, Davidson and Britten later (1979) postulated that the vast majority of structural genes are transcribed continuously at similar rates in any one cell, while the quantitative and qualitative structure of cytoplasmic messenger RNA populations is regulated posttranscriptionally. The basic 'constitutive transcription unit' is assumed to contain a structural gene (including intervening and leader sequences if present) and short interspersed control sequences. Other regions of the genome, which are transcribed in a cell-specific manner upon cell-specific stimuli, do not contain structural genes, but consist of repetitive elements either interspersed with unique spacer sequences or occurring in a clustered arrangement (see also Davidson et al., 1977). Their transcripts are complementary to the short interspersed sequences flanking the structural gene transcript. RNA–RNA duplexes are formed as the result of intranuclear association. These duplexes are thought to protect the structural gene message from a degradative nuclease or to promote messenger RNA processing.

Both the earlier and the expanted version of the model ascribe a regulatory function to the short interspersed middle repetitive sequences. Evidence proposed to be in support of this view includes the following findings:

1) The interspersion of short repetitive sequences with unique sequences has been markedly conserved in the evolution of the eukaryotes (Fig. 4). This genome pattern appears to be typical even for protists. Davidson et al (1975a) have interpreted the prevalence of this pattern to mean that short-period sequence organization provided part of the basis for the evolution of multicellular forms.

2) Short intermediately repetitive sequences are preferentially located in the vicinity of structural genes. This was shown in an analysis of the genome of the sea urchin *Strongylocentrotus purpuratus* (Davidson et al., 1975b). The authors isolated a fraction of DNA in which single-copy sequences were contiguous with

interspersed repetitive sequences. These single-copy sequences represented about one-third of the total unique genomic DNA. Hybridization of the isolated repeat contiguous DNA fraction with messenger RNA of the gastrula stage showed, however, that 80%—100% of the messenger RNA molecules are transcribed from single-copy sequences adjacent to interspersed repeat elements. Similar findings have been made in the exploration of sequence representation in mouse RNA (Kuroiwa and Natori, 1979). In mouse liver and brain, only about half the steady-state nuclear RNA is transcribed from repeat contiguous single-copy sequences, but cytoplasmic RNA is derived mainly (73%—96%) from single-copy sequences adjacent to middle repetitive elements.

In the duck genome a repetitive sequence (or sequences) with a repetitive frequency of about 20 is (are) located in the close vicinity of (some of) the hemoglobin genes (Bishop and Freeman, 1974). The α and β major globin genes of the mouse (Leder et al., 1978) and the histone genes of the sea urchin (Busslinger et al., 1979) also carry short regions of homology in close proximity (see below). Interestingly, in all these cases, the repeat region is located in the 3′ direction of the gene(s), i.e., distal to the initiation of transcription.

Moreover, as Jagodzinski et al. (1979) have shown, middle repetitive sequences of rat DNA, which bind preferentially to nonhistone proteins, are linked to single-copy sequences enriched with regard to representation in liver polysomal RNA. This also would indicate a possible functional significance of the repeat elements concerned.

3) Repetitive sequences are transcribed in a tissue-specific manner. This was shown in the studies of Scheller et al. (1978) and Costantini et al. (1978), investigating the representation of nine, cloned, intermediately repetitive sequences of sea urchin DNA in total oocyte RNA and nuclear RNA of the gastrula stage of embryonic development and of adult intestine cells. It was shown that transcripts of all nine repeat sequences are to be found in all three RNAs, but that the concentration of transcripts of certain repeats differs by more than two orders of magnitude and that repeat-sequence transcripts prevalent in one RNA may be rare in another. It appears that quantitative patterns of repetitive-sequence representation in RNA are cell type-specific. It is important to note that usually, but not always, both DNA strands of the repeat sequence are roughly equally represented in RNA; transcription of both strands is the prerequisite for the formation of RNA–RNA duplexes (see above). It should also be noted that the repeat sequences are represented in polysomal messenger RNA only to a very slight extent, and it cannot be excluded that even this low level may, in fact, be artifactual. This is in accordance with earlier studies of sea urchin embryos (Goldberg et al., 1973), HeLa (Klein et al., 1974), rat (Campo and Bishop, 1974), and mouse (Rabbits and Milstein, 1975; Legler and Cohen, 1976) cells, in which it was shown that transcripts of repetitive sequences are not linked to transcripts of single-copy sequences in polysomal RNA. Conflicting results have been reported from *Xenopus* embryos (Dina et al., 1973, 1974) and *Dictyostelium* (Kindle and Firtel, 1978, 1979; Kimmel and Firtel, 1979). In the last organism several alternative types of sequence organization for transcription units seem to co-exist.

Whereas repetitive sequence representation in nuclear RNA appears to be strongly tissue-specific, this does not seem to be the case with single-copy DNA transcripts. Kleene and Humphreys (1977) have shown that both blastula and pluteus stages of sea urchin *Tripneustes gratilla* embryos transcribe essentially the same sequences, which represent about one-third of the single-copy genome. On the other hand, messenger RNA complexity drastically changes during this developmental period, as reported by Galau et al. (1976b). Similarly, Wold et al. (1978) demonstrated that virtually all sea urchin blastula messenger RNA sequences are present in nuclear RNA of adult intestine cells and adult celomocytes, although most of these messenger RNA sequences are not present in polysomal RNA of these cells. Since the possibility that all nuclear RNAs of diverse sea urchin tissues might contain identical single-copy nuclear RNA sequence sets has been ruled out (Ernst et al., 1979), the cell-specific single-copy transcripts probably do not represent transcripts of structural genes. All these findings are consistent with the model, which does not require coordinate regulation of nuclear and messenger RNA sequences if the majority of the first is not precursor to the latter, or if messenger RNA accumulation is regulated posttranscriptionally.

A number of findings, however, are difficult to reconcile with the proposed general regulatory function of short interspersed repetitive sequences:

1) The short-period sequence-interspersion pattern is common to eukaryote genomes, but it is not at all universal. As pointed out above (Fig. 4), species of diverse lineages of the animal kingdom, and of lower and higher plants display a genomic sequence arrangement typified by longer single-copy and repetitive stretches, lacking the short, 0.2—0.4 kb interspersed repeat class. It might be argued, however, that even in genomes typified by long-period interspersion, structural genes, which often constitute only a small proportion of the genome, are flanked by short repetitive elements undetected with the method used. Unless human and hamster genomes, for example, are functionally organ-

ized in an entirely different manner, this argument would imply that at least most of the short repetitive elements in the *Xenopus*-like genomes are not functionally involved.

2) Genomes exist that consist essentially of non-repeated and interspersed highly repeated sequences only. In such genomes (of true crabs) there are only negligible amounts of lowly and intermediately repetitive DNA (Christie and Skinner, 1979). That messenger RNAs can derive from DNA sequences adjacent to satellite DNA has been documented for one of the *copia* genes in *Drosophila melanogaster* (Carlson and Brutlag, 1978). In addition, the close linkage of short repeated elements with even much shorter single-copy sequences observed in some bird genomes (Ginatulin and Ginatulina, 1979, and personal communication) is difficult to understand in terms of the functional model described above.

3) In the genome of the water mold *Achlya bisexualis,* which is organized in a long-period interspersion pattern, there are approximately 250 interspersed repetitive sequences (Hudspeth et al., 1977). In vegetatively growing cells, however, approximately 2000 single-copy sequences are expressed in messenger RNA. Thus, the number of repeated sequences contiguous with single-copy DNA differs from the number of structural genes expressed at only one developmental stage, by almost one order of magnitude. The authors felt that these observations make it difficult to propose a regulatory function for repeated DNA in this eukaryote.

4) Involvement of short repetitive sequences in regulatory function is expected to impose constraints on changes of the primary nucleotide sequence. However, on comparing the amount of nucleotide substitution in repetitive DNA of related sea urchin species, Chaudhari and Craig (1979) observed no differences between long and short sequences. These authors found that the amount of base-sequence divergence between hybridizable sequences (measured by the depression of the melting point of interspecific DNA hybrids) and the complete loss of hybridizable sequences are no different in long and short repetitive sequences. They suggest that entire families of sequences may be added or deleted at specific phylogenetic points, as already postulated by Mizuno and Macgregor (1974), Flavell and Smith (1976), Flavell et al. (1977), and Moore et al. (1978) on the basis of investigations of a variety of plant and animal model systems. Flavell et al. (1977) have recently discussed possible mechanisms involved in the establishment of the species-specific spectrum of repeated sequences. A certain chromosomal region may be amplified, for example, through some kind of replication error; this event may be followed by cycles of sequence divergence and further amplification. Unequal crossing-over of diverged sequences may play an important role.

It has often been shown that within the same genome, long reassociated repeat sequences display a high thermal stability similar to that of perfectly paired DNA, while the short repetitive sequences show a broad range of thermal stabilities due to sequence mismatch during reassociation (Davidson et al., 1973; Goldberg et al., 1975; Epplen et al., 1979). Britten et al. (1976) have interpreted this observation in the following way: The long repeats, which are more precisely repeated, are relatively recent additions to the genome; they are not necessarily totally new, but they could well be recent multiple copies of pre-existing sequences. [Gillespie (1977) has recently provided experimental evidence in support of this hypothesis, showing a strong phylogenetic specificity of repeated DNA sequences in primates.] The divergence of the short repeated sequences, on the other hand, shows that they have been present for a long period of evolution. The authors assume that these sequences diverge with the same rate as the bulk of single-copy DNA, and that this rate is probably not affected by selection pressure. This could mean that the nucleotide sequence at least of the majority of short repeat DNA elements is not important for regulation. Some short repetitive sequences, however, are highly homologous across species borders [documented, for example in several cereal genomes (Flavell et al., 1977)]. This could well be due to selection because of functional meaning.

Present knowledge of DNA sequence organization allows only indirect conclusions on the mechanism underlying differential gene expression in the eukaryotic cell. The experimental design of the 'statistical' method of studying the functional organization of entire higher genomes is comparatively crude. Most of the work concerns a more or less stringently defined group of sequences, and not sequences with individually specified roles. The development of recombinant DNA techniques, however, which permit clonal replication of eukaryotic DNA segments in bacteria, has brought a rapidly increasing insight into the actual organization of DNA at the gene level. The following sections will give a brief description of the elements supposed to compose typical eukaryotic structural genes and their transcripts, namely:

1) Eukaryotic promoters
2) Leader sequences
3) Exons, introns, and flanking sequences
4) Downstream sequences of homology
5) Messenger RNA caps and ribosome-binding sites
6) Poly(A) tails of messenger RNA.

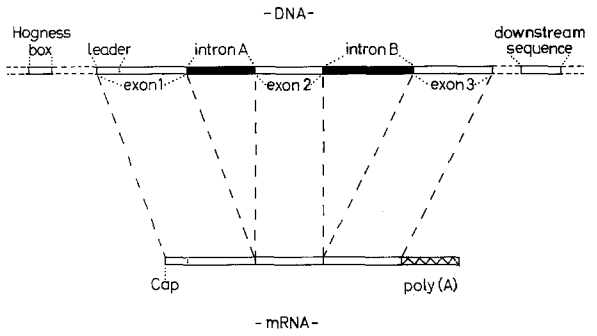

Fig. 5. Schematic presentation of a eukaryotic model gene. The constituents of the structure at the DNA level and at the level of the messenger RNA are described in the text. The mature mRNA is generated most probably in the following scheme: A precursor RNA that is colinear with the gene is transcribed and a poly(A) sequence is added posttranscriptionally. The precursor RNA molecule is then processed to mature RNA by cleavage and removal of intervening and 5′ transcribed sequences

Emphasis is laid on the components that may constitute regulatory elements for the transcription and full expression of a gene.

A eukaryotic model gene is depicted in Fig. 5.

Eukaryotic Promoters

A promoter is functionally defined as a DNA sequence at the 5′ end of a transcription unit, serving as the starting point for RNA polymerase. A promoter *sequence* is not universally defined. Usually, prokaryotic promoters consist of two specific interaction sites: the *recognition site* (Gilbert, 1976) and the '*Pribnow box*' (Pribnow, 1975). The recognition site is centered some 35 base pairs upstream of the first nucleotide coding for messenger RNA, whereas the Pribnow box precedes the initiation of transcription by about ten nucleotides (Scherer et al., 1978; Johnsrud, 1978). However, it is still not known what kind of specific contacts are formed between the RNA polymerase complex and the DNA of the promoter.

Promoters of eukaryotic genes are expected in the same location as in prokaryotes. In fact, analogous to the prokaryotic recognition site, some 30 nucleotides upstream of several eukaryotic genes, a common sequence of six to seven nucleotides was found; it has the form: 5′-TATAAA(T)-3′ (the so-called 'Hogness box' cited by Gannon et al., 1979; Kedes, 1979). This sequence has been found close to the genes coding for sea urchin histone (Schaffner et al., 1978), chick ovalbumin (Gannon et al., 1979), mouse α- and mouse and rabbit major β-globin (Konkel et al., 1978; Nishioka and Leder, 1979), silkworm fibroin (Tsujimoto and Suzuki, 1979a and b), rat insulin (Cordell et al., 1979), and adenovirus-2 major late protein (Ziff and Evans, 1978). Although located in the position of the prokaryotic *recognition site*, these nucleotide sequences (Hogness boxes) closely resemble the prokaryotic Pribnow box. Whether the Hogness box is in fact a common phenomenon of eukaryotic gene organization remains to be established by sequencing of further eukaryotic genes. It has already been shown, however, that eukaryotic genes can be dissimilar with regard to the precise localization of Hogness box-like sequences. In the gene coding for the major protein of hepatitis B virus surface antigen, Valenzuela et al. (1979) were unable to detect this sequence in the close proximity of the nucleotides coding for the *N*-terminal methionine. However, a more distantly related sequence (Ziff and Evans, 1978) occurs about 200 bases away from the 5′ terminus of the adenovirus major late transcript. A similar situation is observed in an IgG 2 light chain (Tonegawa et al., 1978). Galibert et al. (1979), who have cloned the entire hepatitis B virus genome (subtype ayw) in *E. coli,* found only one 'typical' TATAAA sequence per genome, but about 40 similar sequences, differing by only one nucleotide, on each DNA strand. It must be emphasized, however, that a promoter *function* of this common sequence is suggested only by analogy with prokaryotic promoter sequences.

Leader Sequences

Leader sequences of eukaryotic genes are defined as untranslated sequences that code for the 5′ end of mature messenger RNA preceding the AUG initiation codon (Breathnach et al., 1978; Gannon et al., 1979).

Leader sequences have been described for several viral messenger RNAs in eukaryotes (Sambrook, 1977; Chambon, 1977). The leaders of the messenger RNAs sequenced so far vary considerably in length; some of these sequences contain short inverted repeats (palindromes) (Baralle and Brownlee, 1978). It is interesting to note that under in vivo conditions most of the leader sequences are dispensible in simian virus 40 genes (Subramanian, 1979). The same is true for eukaryotic genes: Kronenberg et al. (1979) demonstrated that the first 23—28 nucleotides of the 5′ noncoding region of the β-globin messenger RNA are not essential for protein synthesis in the wheat germ cell-free translational system.

Joining of leader sequences encoded at one location in viral genomes to messenger RNA representing another region of the genome has recently been described in adenovirus, simian virus 40, and avian sarcoma virus systems (Sambrook, 1977; Chow et al., 1977; Krzyzek et al., 1978). This type of RNA processing has been called 'splicing,' i.e., removal of parts of a precursor RNA and rejoining of the ends to form the

mature messenger RNA molecule. Splicing of leader sequences may indeed be a much more general phenomenon, since leader and protein-coding messenger RNA sequences are also spaced by an intervening sequence (see below) and joined in mature messenger RNA in the ovalbumin gene (Gannon et al., 1979). Irrespective of the function of these leader sequences in eukaryotic gene expression, their identification in several virus systems suggests that this kind of splicing phenomenon may be a common feature characteristic of the joint production and/or processing of functionally related messenger RNA species in eukaryotic cells.

Exons, Introns, and Flanking Sequences

An increasing number of experiments show that genes in eukaryotes and their viruses appear to be split: stretches of DNA not represented in mature messenger RNA, the introns, interrupt the coding parts of the gene, the exons (Gilbert, 1978). Bacteria lack intervening DNA sequences. In eukaryotes, the presence of introns is documented in such diverse genes as those coding for ribosomal RNA (Glover and Hogness, 1977; White and Hogness, 1977; Wellauer and Dawid, 1977; Pellegrini et al., 1977; Wellauer et al., 1978; Long and Dawid, 1979); globin (see Kinniburgh and Ross, 1979; Dodgson et al., 1979; Nishioka and Leder, 1979; Konkel et al., 1979; van Ooyen et al., 1979); the chick egg-white proteins ovalbumin (see Royal et al., 1979), ovomucoid (Catterall et al., 1979; Nordstrom et al., 1979), lysozyme (Baldacci et al., 1979), and conalbumin (Cochet et al., 1979; Perrin et al., 1979); rat serum albumin (Sargent et al., 1979); rat growth hormone (Soreq et al., 1979); rat insulin (Lomedico et al., 1979); fibroin of the silk moth (Tsujimoto and Suzuki, 1979a and b); the immunoglobulins (see Abelson, 1979; Honjo et al., 1979); the α-fetoprotein gene of mouse (Tilghman et al., 1979); and human growth hormone and chorionic somatomammotropin (Fiddes et al., 1979). An interesting case is seen in the insertions in ribosomal DNA of ascomycete mitochondria (Bos et al., 1978; Hahn et al., 1979): In view of the hypothesis of Margulis (1970) that mitochondria originated from 'trapped' prokaryotes at an early stage of eukaryote evolution, modern prokaryote genomes may be understood as being 'streamlined' (i.e., freed from intervening sequences) in the course of evolution (Doolittle, 1978), while these mitochondria resemble an ancient, primitive prokaryotic stage. The coding regions of mitochondrial RNAs of *Xenopus laevis* lack intervening sequences (Rastl and Dawid, 1979). If this is a general feature of animal mitochondrial genomes, the similarity in genome structure of the modern bacteria and of the mitochondrial genomes in the more advanced organisms may be regarded as an interesting example of co-evolution.

Some nuclear transfer RNA genes in yeast contain very small intervening sequences of 14—19 base pairs (Goodman et al., 1977; Valenzuela, 1978; Abelson, 1979). Because of the extreme length difference compared with other introns, those of transfer RNAs may stem from a different origin and perhaps have a different significance, if any.

In view of the suggested functional role, it would be of interest to know whether an entire intervening sequence occurs repeatedly in a genome. Very little work has been done to test this directly. Woo et al. (1978) and Mandel et al. (1978) found that the intervening sequences in the chick ovalbumin gene are unique. Recently, Cochet et al. (1979) have shown that high- and low-frequency repetitive sequences are located both upstream of the chicken conalbumin gene and within one intron. Since this is the only known example, further information is required before any conclusions on the matter are possible. It was observed, however, that many introns share short common parts, namely the dinucleotide GT at the 5′, and the dinucleotide AG at the 3′ end (Breathnach et al., 1978). These similarities suggest that the number of different splicing enzymes may be small (Crick, 1979). Again, the introns of the transfer RNAs do not fit this 'rule,' supporting the idea that these intervening sequences evolved differently. There are model systems that appear suitable for providing some insight into the possible functional task of intervening sequences. One of these is the ribosomal DNA of *Drosophila.* In this system, ribosomal genes with and without introns (inside the repeating unit) co-exist. Those that bear introns appear not to be transcribed, whereas the uninterrupted ones do seem to be (Glätzer, 1979; Long and Dawid, 1979).

In the case of immunoglobulins, excision of at least a part of an intron occurs to generate, e.g., the light-chain gene during the peculiar somatic recombination event that takes place in the differentiating mouse lymphocyte (Brack et al., 1978). The resulting gene is characterized by two intervening sequences. The assumed recognition site for a splicing enzyme is marked by a sequence that could form an inverted stem structure (Sakano et al., 1979). These results show that recombination events on both the DNA and the RNA level are involved in expressing an immunoglobulin gene. In contrast, the structure of the fibroin gene is identical in actively expressing and nonexpressing cells as far as the 5′ flanking and intervening sequence (which have been sequenced) are concerned (Tsujimoto and Suzuki, 1979a and b). Similarly, in rat growth hormone (Soreq et al., 1979) and all globin genes of mammals and birds so far described (Lawn et al., 1978; Tilghman et al., 1978; Konkel et al., 1978; van den Berg

et al., 1978; Dodgson et al., 1979) the occurrence and positions of introns appear to be conserved regardless of the cellular source of DNA. A computer-aided comparison of all available globin-gene sequences, including introns, failed to detect fetal-specific sequences (and possibly promoting specific secondary structures) (Smithies et al., 1978). Therefore, differential expression of globin genes during development cannot be explained by changes of occurrence and location of intronic sequences. Although some peculiarities of the split organization of eukaryotic genes may be relevant in the present discussion of gene expression control (see Darnell, 1978; Crick, 1979), any possible conclusions must remain highly speculative for the moment: At least introns cannot have a general function, because uninterrupted eukaryotic genes exist, namely, the clustered sequences coding for histone in *Drosophila* (Hogness, cited according to Crick, 1979) and sea urchins (Schaffner et al., 1978; Levy et al., 1979; Cohn and Kedes, 1979). A sequenced portion of the H2B gene in the sea urchin *Strongylocentrotus purpuratus* is colinear with the 5′ end of the messenger RNA (the nontranslated leader sequence) and also does not contain any insertions (Levy et al., 1978). But there was another peculiarity found in histone genes of *S. purpuratus*: The first transcribed nucleotides map next to a short nucleotide sequence, which is common to all sequenced *S. purpuratus* histone genes (Sures et al., 1976, 1978): 5′ PyCATTCPu-3′; this sequence is found 50—70 bases upstream of the initiation codon for H2A, H2B, H3 and H4. The functional aspect of these homologous sequences remains to be determined.

A strong argument against a functional role of some introns derives from interspecific comparisons of intron nucleotide sequences. For example, when comparing β-globin gene sequences of rabbit and mouse, van Ooyen et al. (1979) observed little sequence homology in the intervening sequences (40%—53%), but a much higher degree of sequence similarity for the coding (81%) and the flanking regions (68%—75%). These findings suggest a functional role for the segments preceding and following the coding regions. Some authors have argued that introns and their splicing have played an important role in eukaryotic gene evolution (Gilbert, 1978; Darnell, 1978). It is possible that exons represent the structural basis of functional 'domains.' Recent studies on immunoglobulin gene structure support this view (Honjo et al., 1979).

Downstream Sequences of Homology

Busslinger et al. (1979) recently investigated the 3′-flanking sequences of all five histone genes of the sea urchin *Psammechinus miliaris*. Some 30 base pairs downstream from the termination codon two highly conserved sequences were found, showing dyad symmetry and therefore suggesting secondary structures as probable recognition signals for regulatory proteins. It is not established whether these sequences are transcribed. Interestingly, the 3′-flanking sequences near *H2A, H2B,* and *H3* genes in *Strongylocentrotus purpuratus* are clearly homologous (Sures et al., 1978) to the respective stretches in the related species, *P. miliaris*. The existence of inverted repeat sequences (or loops) is also shown in the bovine corticotropin-β-lipotropin precursor (Nakanishi et al., 1979) and in the chicken β-globin genes (Dodgson et al., 1979). In mouse α- and major β-globin genes, Leder et al. (1978) demonstrated the conservation of a 3′-flanking sequence 0.15 kb in length, about 1.5 kb away from each gene. In view of the required coordinate regulation of α- and major β-globin genes, which reside on different chromosomes, the finding of such an homologous sequence adjacent to the 3′ end of both coding sequences is remarkable. The regulatory significance, however, remains to be determined.

A further commonly observed flanking element in the 3′-noncoding region of viral and eukaryotic messenger RNA is the sequence 5′-AAUAAA-3′ (Proudfoot and Brownlee, 1976), which usually lies immediately adjacent to the poly(A) tail (Proudfoot and Brownlee, 1974a and b). This evolutionary conserved hexanucleotide has been regarded as a signal for an enzyme, involved either in the synthesis or processing of messenger RNA from its presumed precursor (Perry, 1976). However, not all polyadenylated messenger RNAs of eukaryotes bear this sequence (see Robertson, 1979).

Messenger RNA Caps and Ribosome-binding Sites

The mechanism limiting initiation of translation to the first AUG triplet may or may not involve messenger RNA–ribosomal RNA interactions such as are observed for prokaryotic systems (Steitz and Jakes, 1975; Hagenbüchle et al., 1978). This is because complementarity between the initiation region of eukaryotic messenger RNA and 18S ribosomal RNA is not always obvious (Hagenbüchle et al., 1978). Therefore, Kozak and Shatkin (1978) have proposed a model in which the 40S ribosomal subunit interacts with the 5′ end of the messenger RNA and moves along until it encounters the first AUG. At this stage, joining with the 60S ribosomal subunit may occur. This process is facilitated by the so-called cap.

Most eukaryotic messenger RNAs bear this modification at their 5′ termini; the cap consists of a methylated guanine, which is joined through a 5′-5′ pyrophosphate linkage to a second nucleotide methylated at its 2′ position (Wei and Moss, 1977; Furuichi et al.,

1975). This type of modification of the messenger RNA is known to occur after transcription. The cap structure has been shown to play an important role in the initiation step of protein synthesis (Shatkin, 1976; Kozak and Shatkin, 1978; Plotch et al., 1979), i.e., to make the priming mechanism (the starting of translation) in (e.g.) β-globin messenger RNA 1,000—2,000 times more effective. Not every messenger RNA molecule of a certain gene need carry an identical cap structure. For example, 60% of the fibroin messenger RNA molecules of *Bombyx mori* (Tsujimoto and Suzuki, 1979a) have a methylated cap structure, while the remaining 40% appear to have an unmethylated but otherwise identical cap. As mentioned above, Kronenberg et al. (1979) showed that the first 23—28 bases of rabbit β-globin messenger RNA are not required for protein synthesis in vitro. Yet these results do not mean that the cap is not involved during translation, since a cap-forming activity has been found in the wheat germ assay used (Kronenberg et al., 1979).

Poly(A) Tails of Messenger RNAs

Eukaryotic and viral messenger RNAs generally contain polyadenylate tails of up to 200 residues at their 3′ termini (Lim and Canellakis, 1970; Gaskill and Kabat, 1971). The available results suggest that the 3′ poly(A) sequences have a function in stabilizing the messenger RNA molecules (see Revel and Groner, 1978). Specific proteins may be bound to the poly(A) (Blobel et al., 1973) and thus hinder exonucleases from degrading the messenger. In the absence of poly(A), the half-lives of common messenger RNAs in *Xenopus* oocytes are significantly decreased (Huez et al., 1974), whereas in cell-free extracts the effect is negligible (Sippel et al., 1974; Williamson et al., 1974). Readdition of poly(A) tails to messenger RNA molecules, however, stabilized their expression in the frog oocyte system (Huez et al., 1975).

As mentioned above, a hexanucleotide sequence has been described in the vicinity of poly(A) segments, which is markedly conserved in evolution (Proudfoot and Brownlee, 1976). This sequence may function as a recognition site for an enzyme which is involved with the addition of the poly(A). But Kronenberg et al. (1979) have shown that the 3′-noncoding sequences are not an absolute requirement for translation. These experiments have been confined to in vitro conditions. The in vivo situation is regarded as being much more complicated.

Not only are the histone genes exceptional with regard to DNA sequence structure, but their messenger RNA is also the only example in eukaryotes lacking the posttranscriptional addition of polyadenylate tails (Adesnik and Darnell, 1972; Greenberg and Perry, 1972). The short messenger RNA molecules of histone genes are formed during the synthesis phase of the cell cycle and appear to exit from the nucleus more rapidly than other messenger RNAs (Schochetman and Perry, 1972). Thus, there are at least two mechanisms for the manufacture and transport of messenger RNA, a non-poly(A) and a poly(A) pathway. The latter seems to be realized much more commonly (Darnell et al., 1973).

The recent knowledge of the organization of DNA at the level of the gene has provided many clues as to which elements of a genetic unit may serve regulatory purposes. But fundamental mechanisms of differentiation, at the molecular level, are far from being understood. Further theoretical and experimental work is required to clarify the meaning of DNA sequence organization and the basic design of gene structure.

Acknowledgements. We are grateful to Drs. Hans Küntzel and Jürgen Rimpau for critical discussions of an earlier version of this manuscript. We would like to thank Dr. Wolfgang Engel for his interest and support of our work, and Drs. Pierre Chambon, A. A. Ginatulin and Philippe Kourilsky for the preliminary communication of their results. Finally, we record our gratitude to Mrs. H. Tuczynski and Mr. J. S. Jackson for editorial assistance and Miss I. Kandt for technical assistance.

This project was supported in part by the Deutsche Forschungsgemeinschaft.

References

Abelson, J.: RNA processing and the intervening sequence problem. Annu. Rev. Biochem. **48**, 1035—1069 (1979)

Adesnik, M., Darnell, J. E.: Biogenesis and characterization of histone messenger RNA in HeLa cells. J. Mol. Biol. **67**, 397—406 (1972)

Allen, J. R., Roberts, T. M., Loeblich, A. R. III, Klotz, L. C.: Characterisation of the DNA from the dinoflagellate *Crypthecodinium cohnii* and implications for nuclear organisation. Cell **6**, 161—169 (1975)

Angerer, R. C., Davidson, E. H., Britten, R. J.: DNA sequence organisation in the mollusc *Aplysia californica*. Cell **6**, 29—39 (1975)

Arthur, R. R., Straus, N. A.: DNA-sequence organisation in the genome of the domestic chicken *(Gallus domesticus)*. Can. J. Biochem. **56**, 257—263 (1978)

Bachmann, K.: Genome size in mammals. Chromosoma **37**, 85—93 (1972)

Baldacci, P., Royal, A., Cami, B., Perrin, F., Krust, A., Garapin, A., Kourilsky, P.: Isolation of the lysozyme gene of chicken. Nucleic Acids Res. **6**, 2667—2681 (1979)

Baldari, C. T., Amaldi, F.: DNA reassociation kinetics in relation to genome size in four amphibian species. Chromosoma **59**, 13—22 (1976)

Baldari, C. T., Amaldi, F.: Length and interspersion of repetitive and non-repetitive DNA sequences in four amphibian species with different genome sizes. Chromosoma **61**, 359—368 (1977)

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

Beauchamp, R. S., Pasternak, J., Straus, N. A.: Characterization of the genome of the free-living nematode *Panagrellus*

silusiae: absense of short period interspersion. Biochemistry **18**, 245—251 (1979)

Berg, J. van den, van Ooyen, A., Mantei, N., Schamböck, A., Grosveld, G., Flavell, R. A., Weissmann, C.: Comparison of cloned rabbit and mouse β-globin genes showing strong evolutionary divergence of two homologous pairs of introns. Nature **276**, 37—44 (1978)

Bishop, J. O., Freeman, K. B.: DNA sequences neighboring the duck hemoglobin genes. Cold Spring Harbor Symp. Quant. Biol. **38**, 707—716 (1974)

Blobel, G.: A protein of molecular weight 78,000 bound to the polyadenylate region of eukaryotic messenger RNAs. Proc. Natl. Acad. Sci. USA **70**, 924—928 (1973)

Bonner, T. I., Brenner, D. J., Neufeld, B. R., Britten, R. J.: Reduction in the rate of DNA reassociation by sequence divergence. J. Mol. Biol. **81**, 123—135 (1973)

Borchsenius, S. N., Belozerskaya, N. A., Merkulova, N. A., Wolfson, V. G., Vorob'ev, V. I.: Genome structure of *Tetrahymena pyriformis*. Chromosoma **69**, 275—289 (1978)

Bos, J. L., Heyting, C., Borst, P., Arnberg, A. C., van Bruggen, E. F. J.: An insert in the single gene for the large ribosomal RNA in yeast mitochondrial DNA. Nature **275**, 336—338 (1978)

Bostock, C. J., Sumner, A. T.: The eukaryotic chromosome, pp. 50—58. Amsterdam, New York, Oxford: North-Holland 1978

Boveri, T.: Über Differenzierung der Zellkerne während der Furchung des Eies von *Ascaris megalocephala*. Anat. Anz. **2**, 688—693 (1887)

Brack, C., Hirama, M., Lenhard-Schuller, R., Tonegawa, S.: A complete immunoglobulin gene is created by somatic recombination. Cell **15**, 1—14 (1978)

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

Breuer, M. E., Pavan, C.: Behaviour of polytene chromosomes of *Rhynchosciara angelae* at different stages of larval development. Chromosoma **7**, 371—383 (1955)

Britten, R. J., Davidson, E. H.: Gene regulation for higher cells: a theory. Science **165**, 349—357 (1969)

Britten, R. J., Davidson, E. H.: Repetitive and non-repetitive DNA sequences and a speculation on the origin of evolutionary novelty. Quart. Rev. Biol. **46**, 111—138 (1971)

Britten, R. J., Kohne, D. E.: Repeated sequences in DNA. Science **161**, 529—540 (1968)

Britten, R. J., Smith, J.: A bovine genome. Carnegie Institute Washington. Yearbook **68**, 378—386 (1970)

Britten, R. J., Graham, D. E., Neufeld, B. R.: Analysis of repeating DNA sequences by reassociation. Methods Enzymol. **29**, 363—418 (1974)

Britten, R. J., Graham, D. E., Eden, F. C., Painchaud, D. M., Davidson, E. H.: Evolutionary divergence and length of repetitive sequences in sea urchin DNA. J. Mol. Evol. **9**, 1—23 (1976)

Brykov, V. A., Vol'fson, V. G., Motrenko, E. G., Vorob'ev, V. I.: Unique and repetitive nucleotide sequences in genome of Echinodermata. I. Organization of nucleotide sequences of differing degrees of multiplicity in DNA of Echinodermata. Molek. Biol. (USSR) **12**, 1299—1312 (1978)

Busslinger, M., Portmann, R., Birnstiel, M. L.: A regulatory sequence near the 3′ end of sea urchin histone genes. Nucleic Acids Res. **6**, 2997—3008 (1979)

Campo, M. S., Bishop, J. O.: Two classes of messenger RNA in cultured rat cells: repetitive sequence transcripts and unique sequence transcripts. J. Mol. Biol. **90**, 649—663 (1974)

Carlson, M., Brutlag, D.: One of the *copia* genes is adjacent to satellite DNA in *Drosophila melanogaster*. Cell **15**, 733—742 (1978)

Catterall, J. F., Stein, J. P., Lai, E. C., Woo, S. L. C., Dugaiczyk, A., Mace, M. L., Means, A. R., O'Malley, B. W.: The chick ovomuvoid gene contains at least six intervening sequences. Nature **278**, 323—327 (1979)

Cavalier-Smith, T.: Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. J. Cell Sci. **34**, 247—278 (1978)

Chamberlin, M. E., Britten, R. J., Davidson, E. H.: Sequence organization in *Xenopus* DNA studied by the electron microscope. J. Mol. Biol. **96**, 317—333 (1975)

Chambon, P.: The molecular biology of the eukaryotic genome is coming of age. Cold Spring Harbor Symp. Quant. Biol. **42**, 1209—1234 (1977)

Chaudhari, N., Craig, S. P.: The evolution of the long and short repetitive DNA sequences in sea urchins. Biochim. Biophys. Acta **562**, 438—452 (1979)

Chow, L. T., Gelinas, R. E., Broker, R., Roberts, R. J.: An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. Cell **12**, 1—8 (1977)

Christie, N. T., Skinner, D. M.: Interspersion of highly repetitive DNA with single copy DNA in the genome of the red crab, Geryon quinquedens. Nucleic Acids Res. **6**, 781—796 (1979)

Cochet, M., Gannon, F., Hen, R., Maroteaux, L., Perrin, F., Chambon, P.: Organisation and sequence studies of the 17-piece chicken conalbumin gene. Nature **282**, 567—574 (1979)

Cohn, R. H., Kedes, L. H.: Nonallelic histone gene clusters of individual sea urchins *(Lytechinus pictus)*: Mapping of homologies in coding and spacer DNA. Cell **18**, 855—864 (1979)

Cooke, H. J., Hindley, J.: Cloning of human satellite III DNA: different components are on different chromosomes. Nucleic Acids Res. **6**, 3177—3197 (1979)

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

Costantini, F. D., Scheller, R. H., Britten, R.J., Davidson, E. H.: Repetitive sequence transcripts in the mature sea urchin oocyte. Cell **15**, 173—187 (1978)

Crain, W. R., Eden, F. C., Pearson, W. R., Davidson, E. H., Britten, R.J.: Absence of short period interspersion of repetitive and non-repetitive sequences in the DNA of *Drosophila melanogaster*. Chromosoma **56**, 309—326 (1976a)

Crain, W. R., Davidson, E. H., Britten, R. J.: Contrasting patterns of DNA sequence arrangement in *Apis mellifera* (honeybee) and *Musca domestica* (housefly). Chromosoma **59**, 1—12 (1976b)

Crick, F.: Split genes and RNA splicing. Science **204**, 262—271 (1979)

Darnell, J. E.: Implications of RNA-RNA splicing in evolution of eukaryotic cells. Science **202**, 1257—1260 (1978)

Darnell, J. E., Jelinek, W.R., Molloy, G. R.: Biogenesis of mRNA: genetic regulation in mammalian cells. Science **181**, 1215—1221 (1973)

Davidson, E. H.: Sequence organization in the genome of *Xenopus laevis*. In: Genetic mechanisms of development, F. H. Ruddle (ed.), pp. 251—268. New York, London: Academic Press 1973

Davidson, E. H., Britten, R. J.: Organization, transcription, and regulation in the animal genome. Quart. Rev. Biol. **48**, 565—613 (1973)

Davidson, E. H., Britten, R. J.: Regulation of gene expression: possible role of repetitive sequences. Science **204**, 1052—1059 (1979)

Davidson, E. H., Hough, B. R., Amenson, C. S., Britten, R.J.: General interspersion of repetitive with non-repetitive sequence elements in the DNA of *Xenopus*. J. Mol. Biol. **77**, 1—23 (1973)

Davidson, E. H., Galau, G. A., Angerer, R. C., Britten, R.J.: Comparative aspects of DNA organization in Metazoa. Chromosoma **51**, 253—259 (1975a)

Davidson, E. H., Hough, B. R., Klein, W. H., Britten, R. J.: Structural genes adjacent to interspersed repetitive DNA sequences. Cell **4**, 217—238 (1975b)

Davidson, E. H., Klein, W. H., Britten, R.J.: Sequence organisation in animal DNA and a speculation on hnRNA as a coordinate regulatory transcript. Dev. Biol **55**, 69—84 (1977)

Deininger, P. L., Schmid, C. W.: An electron microscope study of the DNA sequence organization of the human genome. J. Mol. Biol. **106**, 773—790 (1976)

Deininger, P. L., Schmid, C. W.: A study of the evolution of repeated DNA sequences in primates and the existence of a new class of repetitive sequences in primates. J. Mol. Biol. **127**, 437—460 (1979)

Dickinson, D. G., Baker, R. F.: Evidence for translocation of DNA sequences during sea urchin embryogenesis. Proc. Natl. Acad. Sci. USA **75**, 5627—5630 (1978)

Dina, D., Crippa, M., Beccari, E.: Hybridisation properties and sequence arrangement in a population of mRNAs. Nature New Biol. **242**, 101—105 (1973)

Dina, D., Meza, I., Crippa, M.: Relative positions of the 'repetitive', 'unique' and poly(A) fragments of mRNA. Nature **248**, 486—490 (1974)

Dodgson, J. B., Strommer, J., Engel, J. D.: Isolation of the chicken β-globin gene and a linked embryonic β-like globin gene from a chicken DNA recombinant library. Cell **17**, 879—887 (1979)

Donald, L. J., Hamerton, J. L.: A summary of the human gene map, 1973—1977. Cytogenet. Cell Genet. **22**, 5—11 (1978)

Doolittle, W. F.: Genes in pieces: were they ever together? Nature **272**, 581—582 (1978)

Dott, P. J., Chuang, C. R., Saunders, G. F.: Inverted repetitive sequences in the human genome. Biochemistry **15**, 4120—4125 (1976)

Eden, F. C., Hendrick, J.P.: Unusual organization of DNA sequences in the chicken. Biochemistry **17**, 5838—5844 (1978)

Eden F. C., Graham, D. E., Davidson, E. H., Britten, R. J.: Exploration of long and short repetitive sequence relationships in the sea urchin genome. Nucleic Acids Res. **4**, 1553—1567 (1977)

Efstratiadis, A., Crain, W. R., Britten, R. J., Davidson, E. H., Kafatos, F. C.: DNA sequence organization in the lepidopteran *Antheraea pernyi*. Proc. Natl. Acad. Sci. USA **73**, 2289—2293 (1976)

Epplen, J.T., Leipoldt, M., Engel, W., Schmidtke, J.: DNA sequence organisation in avian genomes. Chromosoma **69**, 307—321 (1978)

Epplen, J. T., Diedrich, U., Wagenmann, M., Schmidtke, J., Engel, W.: Contrasting DNA sequence organisation patterns in sauropsidian genomes. Chromosoma **75**, 199—214 (1979)

Ernst, S. G., Britten, R. J., Davidson, E. H.: Distinct single-copy sequence sets in sea urchin nuclear RNAs. Proc. Natl. Acad. Sci. USA **76**, 2209—2212 (1979)

Fiddes, J. C., Seeburg, P. H., DeNoto, F. M., Hallewell, R. A., Baxter, J. D., Goodman, H. M.: Structure of genes for human growth hormone and chorionic somatomammotropin. Proc. Natl. Acad. Sci. USA **76**, 4294—4298 (1979)

Firtel, R. A., Kindle, K.: Structural organisation of the genome of the cellular slime mold *Dictyostelium discoideum*: Interspersion of repetitive and single-copy DNA sequences. Cell **5**, 401—411 (1975)

Flavell, R. B., Smith, D. B.: Nucleotide sequence organisation in the wheat genome. Heredity **37**, 231—252 (1976)

Flavell, R. B., Rimpau, J., Smith, D. B.: Repeated sequence DNA relationships in four cereal genomes. Chromosoma **63**, 205—222 (1977)

Furuichi, Y., Morgan, M., Muthukrishnan, S., Shatkin, A. J.: Reovirus messenger RNA contains a methylated, blocked 5′-terminal structure: m^7 G(5′)ppp(5′)G^m pCp-. Proc. Natl. Acad. Sci. USA **72**, 362—366 (1975)

Galau, G. A., Chamberlin, M. E., Hough, B. R., Britten, R. J., Davidson, E. H.: Evolution of repetitive and non-repetitive DNA. In: Molecular evolution. Ayala, F. J. (ed.), pp. 200—224. Sunderland, Mass.: Sinauer Ass. 1976a

Galau, G. A., Klein, W. H., Davis, M. M., Wold, B. J., Britten, R. J., Davidson, E. H.: Structural gene sets active in embryos and adult tissues of the sea urchin. Cell **7**, 487—505 (1976b)

Galibert, F., Mandart, E., Fitoussi, F., Tiollais, P., Charnay, P.: Nucleotide sequence of the hepatitis B virus genome (subtype ayw) cloned in *E. coli*. Nature **281**, 646—650 (1979)

Gannon, F., O'Hare, K., Perrin, F., LePennec, J. P., Benoist, C., Cochet, M., Breathnach, R., Royal, A., Garapin, A., Cami, B., Chambon, P.: Organisation and sequences at the 5′ end of a cloned complete ovalbumin gene. Nature **278**, 428—434 (1979)

Gaskill, P., Kabat, D.: Unexpectedly large size of globin messenger ribonucleic acid. Proc. Natl. Acad. Sci. USA **68**, 72—75 (1971)

Gilbert, W.: Starting and stopping sequences for the RNA polymerase. In: RNA polymerase. Losick, R., Chamberlin, M. (eds.), pp. 193—205. New York: Cold Spring Harbour Laboratory 1976

Gilbert, W.: Why genes in pieces? Nature **271**, 501 (1978)

Gillespie, D.: Newly evolved repeated DNA sequences in primates. Science **196**, 889—891 (1977)

Ginatulin, A. A., Ginatulina, L. K.: Molecular structure of genome of vertebrates. In: Evolutionary studies. Parallelism and convergence (Krassilov, V. A., ed.). Vladivostok 1979

Ginelli, E., Corneo, G.: The organization of repeated DNA sequences in the human genome. Chromosoma **56**, 55—68 (1976)

Ginelli, E., Di Lernia, R., Corneo, G.: The organization of DNA sequences in the mouse genome. Chromosoma **61**, 215—226 (1977)

Glätzer, K. H.: Lengths of transcribed rDNA repeating units in spermatocytes of *Drosophila hydei*: only genes without an intervening sequence are expressed. Chromosoma **75**, 161—175 (1979)

Glover, D. M., Hogness, D. S.: A novel arrangement of the 18S and 28S sequences in a repeating unit of *Drosophila melanogaster* rDNA. Cell **10**, 167—176 (1977)

Goldberg, R. B., Galau, G. A., Britten, R. J., Davidson, E. H.: Nonrepetitive DNA sequence representation in sea urchin embryo messenger RNA. Proc. Natl. Acad. Sci. USA **70**, 3516—3520 (1973)

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., Davidson, E. H.: DNA sequence organization in the genomes of five marine invertebrates. Chromosoma **51**, 225—251 (1975)

Goodman, H. M., Olson, M. V., Hall, B. D.: Nucleotide sequence of a mutant eukaryotic gene: the yeast tyrosine-inserting ochre suppressor SUP4-o. Proc. Natl. Acad. Sci. USA **74**, 5453—5457 (1977)

Graham, D. E., Neufeld, B. R., Davidson, E. H., Britten, R. J.: Interspersion of repetitive and non-repetitive DNA sequences in the sea urchin genome. Cell **1**, 127—137 (1974)

Greenberg, J. R., Perry, R. P.: Relative occurrence of polyadenylic acid sequences in messenger and heterogenous nuclear RNA of L cells as determined by poly(U)-hydroxylapatite chromatography. J. Mol. Biol. **72**, 91—98 (1972)

Hagenbüchle, O., Santer, M., Steitz, J. A., Mans, R. J.: Conservation of the primary structure at the 3′ end of 18S rRNA from eukaryotic cells. Cell **13**, 551—563 (1978)

Hahn, U., Lazarus, C. M., Lünsdorf, H., Küntzel, H.: Split gene for mitochondrial 24S ribosomal RNA of *Neurospora crassa*. Cell **17**, 191—200 (1979)

Harshey, R. M., Jayaram, M., Chamberlin, M. E.: DNA sequence organization in *Phycomyces blakesleeanus*. Chromosoma **73**, 143—151 (1979)

Herzfeld, F., Kiper, M.: Organization and function of the eukaryotic genome. Progr. Bot. **41**, 198—211 (1979)

Hinegardner, R.: Evolution of genome size. In: Molecular evolution. Ayala, F. J. (ed.), pp. 179—199. Sunderland, Mass.: Sinauer Ass. 1976

Honjo, T., Obata, M., Yamawaki-Kataoka, Y., Kataoka, T., Kawakami, T., Takahashi, N., Mano, Y.: Cloning and complete nucleotide sequence of mouse immunoglobulin γ1 chain gene. Cell **18**, 559—568 (1979)

Houck, C. M., Rinehart, F. P., Schmid, C. W.: Fractionation of renatured repetitive human DNA according to thermal stability, sequence length, and renaturation rate. Biochim. Biophys. Acta **518**, 37—52 (1978)

Hudspeth, M. E. S., Timberlake, W. E., Goldberg, R. B.: DNA sequence organization in the water mold *Achlya*. Proc. Natl. Acad. Sci. USA **74**, 4332—4336 (1977)

Huez, G., Marbaix, G., Hubert, E., Leclercq, M., Nudel, U., Soreq, H., Salomon, R., Lebleu, B., Revel, M., Littauer, U. Z.: Role of the polyadenylate segment in the translation of globin messenger RNA in *Xenopus* oocytes. Proc. Natl. Acad. Sci. USA **71**, 3143—3146 (1974)

Huez, G., Marbaix, G., Hubert, E., Cleuter, Y., Leclercq, M., Chantrenne, H., Devos, R., Soreq, H., Nudel, U., Littauer, U. Z.: Readenylation of polyadenylate-free globin messenger RNA restores its stability in vivo. Eur. J. Biochem. **59**, 589—592 (1975)

Jacob, F., Monod, J.: Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. **3**, 318—356 (1961)

Jagodzinski, L. L., Castro, C. E., Sherrod, P., Lee, D., Sevall, J. S.: Reassociation kinetics of non-histone bound DNA sites. J. Biol. Chem. **254**, 3038—3044 (1979)

Johnsrud, L.: Contacts between *Escherichia coli* RNA polymerase and a *lac* operon promoter. Proc. Natl. Acad. Sci. USA **75**, 5314—5318 (1978)

Jones, K. W.: Speculations on the functions of satellite DNA in evolution. Z. Morphol. Anthropol. **69**, 143—171 (1978)

Kedes, L. H.: Histone genes and histone messengers. Annu. Rev. Biochem. **48**, 837—870 (1979)

Kimmel, A., Firtel, R. A.: A family of short, interspersed repeat sequences at the 5′ end of a set of *Dictyostelium* single-copy mRNAs. Cell **16**, 787—796 (1979)

Kindle, K. L., Firtel, R. A.: Identification and analysis of *Dictyostelium* actin genes, a family of moderately repeated genes. Cell **15**, 763—778 (1978)

Kindle, K. L., Firtel, R. A.: Evidence that populations of *Dictyostelium* single-copy mRNA transcripts carry common repeat sequences. Nucleic Acids Res. **6**, 2403—2422 (1979)

King, M.-C., Wilson, A. C.: Evolution at two levels in humans and chimpanzees. Science **188**, 107—116 (1975)

Kinniburgh, A. J., Ross, J.: Processing of the mouse β-globin mRNA precursor: at least two cleavage-ligation reactions are necessary to excise the larger intervening sequence. Cell **17**, 915—921 (1979)

Kiper, M.: Gene numbers as measured by single-copy DNA saturation with mRNA are routinely overestimates. Nature **278**, 279—280 (1979)

Kleene, K. C., Humphreys, T.: Similarity of hnRNA sequences in blastula and pluteus stage sea urchin embryo. Cell **12**, 143—155 (1977)

Klein, W. H., Murphy, W., Attardi, G., Britten, R. J., Davidson, E. H.: Distribution of repetitive and non-repetitive sequence transcripts in HeLa mRNA. Proc. Natl. Acad. Sci. USA **71**, 1785—1789 (1974)

Konkel, D. A., Tilghman, S. M., Leder, P.: The sequences of the chromosomal mouse β-globulin major gene: Homologies in capping, splicing and poly(A) sites. Cell **15**, 1125—1132 (1978)

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

Kozak, M., Shatkin, A. J.: Identification of features in 5′ terminal fragments from reovirus mRNA which are important for ribosome binding. Cell **13**, 201—212 (1978)

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

Krzyzek, R. A., Collett, M. S., Lau, A. F., Pardue, M. L., Leis, J. P., Faras, A. J.: Evidence for splicing of avian sarcoma virus 5′-terminal genomic sequences onto viral-specific RNA in infected cells. Proc. Natl. Acad. Sci. USA **75**, 1284—1288 (1978)

Kuroiwa, A., Natori, S.: Preferential expression of unique sequences adjacent to middle repetitive sequences in mouse cytoplasmic RNA. Nucleic Acids Res. **7**, 751—764 (1979)

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

Leder, A., Miller, H. I., Hamer, D. H., Seidman, J. G., Norman, B., Sullivan, M., Leder, P.: Comparison of cloned mouse α- and β-globin genes: Conservation of intervening sequence locations and extragenic homology. Proc. Natl. Acad. Sci. USA **75**, 6187—6191 (1978)

Lee, A. S., Britten, R. J., Davidson, E. H.: Interspersion of short repetitive sequences studied in cloned sea urchin DNA fragments. Science **196**, 189—192 (1977)

Legler, M. K., Cohen, E. P.: Estimation of the number of nucleotide sequences in mouse DNA complementary to messenger RNAs specifying a complete mouse immunoglobulin. Biochemistry **15**, 4390—4399 (1976)

Levy, S., Sures, I., Kedes, L. H.: Sequence of the 5′-end of *Strongylocentrotus purpuratus* H2b histone mRNA and its location within histone DNA. Nature **279**, 737—739 (1979)

Lim, L., Canellakis, E. S.: Adenine-rich polymer associated with rabbit reticulocyte messenger RNA. Nature **227**, 710—712 (1970)

Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R., Tizard, R.: The structure and evolution of the two nonallelic rat preproinsulin genes. Cell **18**, 545—558 (1979)

Long, E. O., Dawid, I. B.: Expression of ribosomal DNA insertion in *Drosophila melanogaster*. Cell **18**, 1185—1196 (1979)

Mandel, J. L., Breathnach, R., Gerlinger, P., Le Meur, M., Gannon, F., Chambon, P.: Organization of coding and intervening sequences in the chicken ovalbumin split gene. Cell **14**, 641—653 (1978)

Manning, J. E., Schmid, C. W., Davidson, N.: Interspersion of repetitive and nonrepetitive DNA sequences in the *Drosophila melanogaster* genome. Cell **4**, 141—155 (1975)

Margulis, L.: Origin of eukaryotic cells. New Haven, London: Yale University Press 1970

Mattern, M. R., Painter, R. B.: The organization of repeated nucleotide sequences in the replicons of mammalian DNA. Biophys. J. **19**, 117—123 (1977)

Miklos, G. L. G., John, B.: Heterochromatin and satellite DNA in man: properties and prospects. Am. J. Hum. Genet. **31**, 264—280 (1979)

Mizuno, S., Macgregor, H. C.: Chromosomes, DNA sequences, and evolution in salamanders of the genus *Plethodon*. Chromosoma **48**, 239—296 (1974)

Moore, G. P., Scheller, R. H., Davidson, E. H., Britten, R. J.: Evolutionary change in the repetition frequency of sea urchin DNA sequences. Cell **15**, 649—660 (1978)

Moritz, K. B., Roth, G. E.: Complexity of germline and somatic DNA in *Ascaris*. Nature **259**, 55—57 (1976)

Moyzis, R., Bonnet, J., Ts'o, P. O. P.: DNA sequence organization of the syrian hamster. J. Cell Biol. **75**, 130a (1977)

Nakanishi, S., Inoue, A., Kita, T., Nakamura, M., Chang, A. C. Y., Cohen, S. N., Numa, S.: Nucleotide sequence of cloned cDNA for bovine corticotropin-β-lipotropin precursor. Nature **278**, 423—427 (1979)

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

Nordstrom, J. L., Roop, D. R., Tsai, M.-J., O'Malley, B. W.: Identification of potential ovomucoid mRNA precursors in chick oviduct nuclei. Nature **278**, 328—331 (1979)

Ohno, S.: Simplicity of mammalian regulatory systems inferred by single gene determination of sex phenotypes. Nature **234**, 134—137 (1971)

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

Pearson, W. R., Wu, J.-R., Bonner, J.: Analysis of rat repetitive RNA sequences. Biochemistry **17**, 51—59 (1978)

Pellegrini, M., Manning, J., Davidson, N.: Sequence arrangement of the rDNA of Drosophila melanogaster. Cell **10**, 213—224 (1977)

Perrin, F., Cochet, M., Gerlinger, P., Cami, B., Le Pennec, J. P., Chambon, P.: The chicken conalbumin gene: studies of the organization of cloned DNAs. Nucleic Acids Res. **6**, 2731—2748 (1979)

Perry, R. P.: Processing of RNA. Annu. Rev. Biochem. **45**, 605—629 (1976)

Plotch, S. J., Bouloy, M., Krug, R. M.: Transfer of 5'-terminal cap of globin mRNA to influenza viral complementary RNA during transcription in vitro. Proc. Natl. Acad. Sci. USA **76**, 1618—1622 (1979)

Pribnow, D.: Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. Proc. Natl. Acad. Sci. USA **72**, 784—788 (1975)

Proudfoot, N. J., Brownlee, G. G.: Nucleotide sequence adjacent to polyadenylic acid in globin messenger RNA. FEBS Lett. **38**, 179—183 (1974a)

Proudfoot, N. J., Brownlee, G. G.: Sequence at the 3' end of globin mRNA shows homology with immunoglobulin light chain mRNA. Nature **252**, 359—362 (1974b)

Proudfoot, N. J., Brownlee, G. G.: 3' Non-coding region sequences in eukaryotic messenger RNA. Nature **263**, 211—214 (1976)

Rabbits, T. H., Milstein, C.: Mouse immunoglobulin genes: studies on the reiteration frequency of light-chain genes by hybridisation procedures. Eur. J. Biochem. **52**, 125—133 (1975)

Rastl, E., Dawid, I. B.: Expression of the mitochondrial genome in *Xenopus laevis*: a map of transcripts. Cell **18**, 501—510 (1979)

Rawson, J. R. Y., Eckenrode, V. K., Boerma, C. L., Curtis, S.: DNA sequence organisation in the alga *Euglena gracilis*. Biochim. Biophys. Acta **563**, 1—16 (1979)

Remane, A., Storch, V., Welsch, U.: Systematische Zoologie. Stuttgart: Fischer 1976

Revel, M., Groner, Y.: Post-transcriptional and translational controls of gene expression in eukaryotes. Annu. Rev. Biochem. **47**, 1079—1126 (1978)

Rimpau, J., Smith, D., Flavell, R.: Sequence organisation of the wheat and rye genomes by interspecies DNA/DNA hybridisation. J. Mol. Biol. **123**, 327—359 (1978)

Robertson, J. S.: 5' and 3' terminal nucleotide sequences of the RNA genome segments of influenza virus. Nucleic Acids Res. **6**, 3745—3757 (1979)

Roth, G. E.: Satellite DNA properties of the germ line limited DNA and the organisation of the somatic genomes in the Nematodes *Ascaris suum* and *Parascaris equorum*. Chromosoma **74**, 355—371 (1979)

Royal, A., Garapin, A., Cami, B., Perrin, F., Mandel, J. L., Le Meur, M., Brégégègre, F., Gannon, F., LePennec, J. P., Chambon, P., Kourilsky, P.: The ovalbumin gene region: common features in the organisation of three genes expressed in chicken oviduct under hormonal control. Nature **279**, 125—132 (1979)

Sakano, H., Hüppi, K., Heinrich, G., Tonegawa, S.: Sequences at the somatic recombination sites of immunoglobulin light-chain genes. Nature **280**, 288—294 (1979)

Sambrook, J.: Adenovirus amazes at Cold Spring Harbor. Nature **268**, 101—104 (1977)

Samols, D., Swift, H.: Genomic organisation in the flesh fly *Sarcophaga bullata*. Chromosoma **75**, 129—143 (1979)

Sargent, T. D., Wu, J.-R., Sala-Trepat, J. M., Wallace, R. B., Reyes, A. A., Bonner, J.: The rat serum albumin gene: Analysis of cloned sequences. Proc. Natl. Acad. Sci. USA **76**, 3256—3260 (1979)

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

Scheller, R. A., Costantini, F. D., Kozlowski, M. R., Britten, R. J., Davidson, E. H.: Specific representation of cloned repetitive DNA sequences in sea urchin RNAs. Cell **15**, 189—203 (1978)

Scherer, G. E. F., Walkinshaw, M. D., Arnott, S.: A computer aided oligonucleotide analysis provides a model sequence for RNA polymerase-promoter recognition in *E. coli*. Nucleic Acids Res. **5**, 3759—3773 (1978)

Schmid, C. W., Deininger, P. L.: Sequence organisation of the human genome. Cell **6**, 345—358 (1975)

Schmid, C. W., Manning, J. E., Davidson, N.: Inverted repeat sequences in the *Drosophila* genome. Cell **5**, 159—172 (1975)

Schmidtke, J., Epplen, J. T., Engel, W.: Genome analysis of *Amphioxus* and speculation as to the origin of contrasting vertebrate genome organisation patterns. Comp. Biochem. Physiol. **63B**, 455—458 (1979a)

Schmidtke, J., Schmitt, E., Leipoldt, M., Engel, W.: Amount of repeated and non-repeated DNA in the genomes of closely related fish species with varying genome sizes. Comp. Biochem. Physiol. **64B**, 117—120 (1979b)

Schmidtke, J., Schmitt, E., Matzke, E., Engel, W.: Non-repetitive DNA sequence divergence in phylogenetically diploid and tetraploid teleostean species of the family *Cyprinidae* and the order *Isospondyli*. Chromosoma **75**, 185—198 (1979c)

Schochetman, G., Perry, R. P.: Early appearance of histone messenger RNA in polyribosomes of cultured L cells. J. Mol. Biol. **63**, 591—596 (1972)

Shatkin, A. J.: Capping of eukaryotic mRNAs. Cell **9**, 645—653 (1976)

Singh, L., Purdom, I. F., Jones, K. W.: Satellite DNA and evolution of sex chromosomes. Chromosoma **59**, 43—62 (1976)

Sippel, A. E., Stavrianopoulos, J. G., Schütz, G., Feigelson, P.: Translational properties of rabbit globin mRNA after specific removal of poly(A) with ribonuclease H. Proc. Natl. Acad. Sci. USA **71**, 4635—4639 (1974)

Smith, H. O.: Nucleotide sequence specificity of restriction endonucleases. Science **205**, 455—462 (1979)

Smith, M. J., Boal, R.: DNA sequence organisation in the common Pacific starfish *Pisaster ochraceous*. Can. J. Biochem. **56**, 1048—1054 (1978)

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

Soreq, H., Harpold, M., Evans, R., Darnell, J. E., Bancroft, F. C.: Rat growth hormone gene: intervening sequences separate the mRNA regions. Nucleic Acids Res. **6**, 2471—2482 (1979)

Steitz, J. A., Jakes, K.: How ribosomes select initiator regions in mRNA: base pair formation between the 3′ terminus of 16S rRNA and the mRNA during initiation of protein synthesis in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **72**, 4734—4738 (1975)

Strom, C. M., Moscona, M., Dorfman, A.: Amplification of DNA sequences during chicken cartilage and neural retina differentiation. Proc. Natl. Acad. Sci. USA **75**, 4451—4454 (1978)

Subramanian, K. N.: Segments of simian virus 40 DNA spanning most of the leader sequence of the major late viral messenger RNA are dispensable. Proc. Natl. Acad. Sci. USA **76**, 2556—2560 (1979)

Sures, I., Maxam, A., Cohn, R. H., Kedes, L. H.: Identification and location of the histone H2A and H3 genes by sequence analysis of sea urchin (*S. purpuratus*) DNA cloned in *E. coli*. Cell **9**, 495—502 (1976)

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

Thompson, W. F.: Perspectives on the evolution of plant DNA. Carnegie Institute Washington Yearbook **77**, 310—316 (1978)

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

Tilghman, S. M., Kioussis, D., Gorin, M. B., Ruiz, J. P. G., Ingram, R. S.: The presence of intervening sequences in the α-fetoprotein gene of the mouse. J. Biol. Chem. **254**, 7393—7399 (1979)

Timberlake, W. E.: Low repetitive DNA content in *Aspergillus nidulans*. Science **202**, 973—975 (1978)

Tonegawa, S., Maxam, A. M., Tizard, R., Bernard, O., Gilbert, W.: Sequence of a mouse germ-line gene for a variable region of an immunoglobulin light chain. Proc. Natl. Acad. Sci. USA **75**, 1485—1489 (1978)

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

Tsujimoto, Y., Suzuki, Y.: The DNA sequence of *Bombyx mori* fibroin gene including the 5′ flanking, mRNA coding, entire intervening and fibroin protein coding regions. Cell **18**, 591—600 (1979b)

Valenzuela, P., Venegas, A., Weinberg, F., Bishop, R., Rutter, W. J.: Structure of yeast phenylalanine-tRNA genes: an intervening DNA segment within the region coding for the tRNA. Proc. Natl. Acad. Sci. USA **75**, 190—194 (1978)

Valenzuela, P., Gray, P., Quiroga, M., Zaldivar, J., Goodman, H. M., Rutter, W. J.: Nucleotide sequence of the gene coding for the major protein of hepatitis B virus surface antigen. Nature **280**, 815—819 (1979)

Vogel, F., Motulsky, A. G.: Human genetics: Problems and approaches. Berlin, Heidelberg, New York: Springer 1979

Wei, C.-M., Moss, B.: 5′-terminal capping of RNA by guanylyltransferase from HeLa cell nuclei. Proc. Natl. Acad. Sci. USA **74**, 3758—3761 (1977)

Wellauer, P. K., Dawid, I. B.: The structural organization of ribosomal DNA in *Drosophila melanogaster*. Cell **10**, 193—212 (1977)

Wells, R., Royer, H.-D., Hollenberg, C. P.: Non-Xenopus-like DNA sequence organisation in the *Chironomus tentans* genome. Mol. Gen. Genet. **147**, 45—51 (1976)

Wensink, P. C., Tabata, S., Pachl, C.: The clustered and scrambled arrangement of moderately repetitive elements in *Drosophila* DNA. Cell **18**, 1231—1246 (1979)

White, R. L., Hogness, D. S.: R loop mapping of the 18S and 28S sequences in the long and short repeating units of *Drosophila melanogaster* rDNA. Cell **10**, 177—192 (1977)

Wilkes, M. M., Pearson, W. R., Wu, J.-R., Bonner, J.: Sequence organisation of the rat genome by electron microscopy. Biochemistry **17**, 60—69 (1978)

Williamson, R., Crossley, J., Humphries, S.: Translation of mouse globin messenger ribonucleic acid from which the poly(adenylic acid) sequence has been removed. Biochemistry **13**, 703—707 (1974)

Wimpee, C. F., Rawson, J. R. Y.: Characterization of the nuclear genome of pearl millet. Biochim. Biophys. Acta **562**, 192—206 (1979)

Wold, B. J., Klein, W. H., Hough-Evans, B. R., Britten, R. J., Davidson, E. H.: Sea urchin embryo mRNA sequences expressed in the nuclear RNA of adult tissues. Cell **14**, 941—950 (1978)

Woo, S. L. C., Dugaiczyk, A., Tsai, M.-J., Lai, E. C., Catterall, J. F., O'Malley, B. W.: The ovalbumin gene: cloning of the natural gene. Proc. Natl. Acad. Sci. USA **75**, 3688—3692 (1978)

Wu, J.-R., Pearson, W. R., Posakony, J. W., Bonner, J.: Sequence relationship between long and short repetitive DNA of the rat: A preliminary report. Proc. Natl. Acad. Sci. USA **74**, 4382—4386 (1977)

Ziff, E. B., Evans, R. M.: Coincidence of the promoter and capped 5′ terminus of RNA from the adenovirus 2 major late transcription unit. Cell **15**, 1463—1475 (1978)

Received January 21, 1980

Note Added in Proof

Preliminary results from our laboratory suggest that the genome of the Syrian hamster is arranged in a short period sequence interspersion pattern (Leipoldt et al., unpublished); this finding is in contrast to the report of Moyzis et al. (1977).