Genetica **115:** 105–129, 2002.

Why so many noncoding nucleotides ? The eukaryote genome as an epigenetic machine

Emile Zuckerkandl
Institute of Molecular Medical Sciences, P.O. Box 20452, Stanford, CA 94309, USA (Phone: (650) 617-1850; E-mail: EmileIMMS@aol.com)

Key words: c-value paradox, cell determination, dispensability of sequences, functional density, gene interaction complexity, introns, position effect variegation, transcriptional regulation

Abstract

It is recalled that dispensability of sequences and neutral substitution rate must not be construed to be markers of nonfunctionality. Different aspects of functionality relate to differently-sized nucleotide communities. At the time cells became nucleated, a boom of epigenetic processes led to uses of DNA that required many more nucleotides operating collectively than do functions definable in terms of classical genetics. Each order of magnitude of nucleotide plurality was colonized by functions germane to that order. The eukaryote genome became a great epigenetic machine. Sequences of different levels of nucleotide plurality are briefly discussed from the point of view of their functional relevance. By their activities as both transcribed genes and *cis*-acting repeats, SINEs and LINEs are the principal link between genetic and epigenetic processes. SINEs can act as local repeats to produce position effect variegation (PEV) in a nearby gene. PEV may thus represent a general method of overall transcriptional regulation at the level of cell collectivities. When tracking the scale dependence of nucleotide function, one finds the 100 kb order of nucleotide plurality to provide epigenetically the basis at once for PEV, imprinting, and cell determination, with sectorial repressibility a trait common to the three. In sectorial repressibility, introns may play a structural role favoring the stability of higher-order chromatin structures. At that level of nucleotide involvement, nonconserved nonhomologous nonprotein-coding sequences may often play the same structural roles. In addition, genomic distance per se – and, therefore, the mass of intervening nucleotides – can have functional effects. Distances between enhancers and promoters need to be probed in this respect. At the 1000 kb level of nucleotide function, attention is focused on the formation of centromeres. It is one of the levels of nucleotide plurality per function where specificity in the generation of DNA/protein complexes seems to depend more upon the structural fit among factors than upon the DNA sequence. This circumstance may explain in part the prevailing difficulty in recognizing the functional nature of sequences among non-protein-coding nucleotide arrays and the propensity among investigators to tag the majority of DNA sequences in higher organisms as functionally meaningless. Noncoding DNA often may not be 'selected' as an appropriate niche for a certain function, but be 'elected' in that capacity by a group of factors, as a preexisting sequence that is only now called upon to serve. Much of the non-protein-coding DNA may thus be only conditionally functional and in fact may never be elected to functions at a high level of nucleotide plurality. Eukaryotes are composites, at different levels of this plurality, of the functional and the nonfunctional, as well as of the conditionally functional and the outright functional. Thus, a sequence that is nonfunctional at one level of nucleotide plurality may participate in a functional sequence at a more inclusive level. In the end, every nucleotide is at least infinitesimally functional if, for metabolic and developmental reasons, the chromatin mass as such becomes a selectable entity. Given the scale dependence of nucleotide function, large amounts of 'junk DNA', contrary to common belief, must be assumed to contribute to the complexity of gene interaction systems and of organisms.

Aims of the present analysis

Genome sizes can differ by five orders of magnitude among eukaryotes (cf. Li & Graur, 1991), and sometimes vary considerably over relatively closely related taxa (Cavalier-Smith, 1978; Petrov et al., 2000). For example, the brown mountain grasshopper *Podisma* has a genome 10 times as large as that of the cricket *Laupala* and over one hundred times as large as that of *Drosophila* (Bensasson et al., 2001). Genome size varies 'only' by a factor of 364 among the vertebrates examined (Cavalier-Smith, 1991). Some of the important mechanisms of this variation have recently been revealed (Petrov & Hartl, 1998; Petrov et al., 2000; Bensasson et al., 2001; Petrov, 2001).

About 95% of haploid genomes in multicellular eukaryotes have been widely considered as 'junk'. The term is picturesque, but not insightful. A number of biologists, among them notably Susumo Ohno (1997), the originator of the notion of junk DNA, clung persistently to this position without the slightest nuancing. 'Junk DNA' stands for DNA that is functionless at a given moment in evolutionary time. It may seem easy to define junk, yet, it is only as easy as it is to define function, which is not easy.

True, there are the easy cases. It happens frequently that a relatively short DNA sequence – say, a few hundred base pairs long – not endowed with any obvious function, is requisitioned for an obvious function. The integration of additional structures into a preexisting structural-functional ensemble takes place in haphazard, makeshift ways – in quirky ways (a phrase used by Arian Smit, 1999). For example, a fragment of an Alu sequence can become part of an apparently functional protein-coding sequence (Makalowski, Mitchell & Labuda, 1994). (Alu sequences do not normally code for polypeptide chains.) Other mobile elements have secondarily been called upon as participants in the transcriptional control of genes (Britten, 1996a, b). Examples of such functional takeovers abound. They are thought to represent a removal of some individual item from a genomic junk yard. The items are not always small, as illustrated by neoformation of centromeres. It is generally thought that if the sequences are *not* mobilized for what must obviously be a function, they will be rapidly changed or lost.

This last statement sounds innocuous, yet is misleading. A sequence requisitioned for a function? Strictly speaking, often (and, in one sense, even generally!) probably for *another* and superimposed function. When not a winner in the lottery game of adoption for a function, the sequence will be rapidly changed? But rapid sequence change does not imply absence of function! The sequence can be lost because it is functionless? It can be lost whether functionless or not!

These qualifications of the statement in question reflect a thesis of the present paper. It is not just the slowness of growth in our knowledge that retards a consensus on functions in genomes; I would contend that the major obstacle is a widespread narrowness in the perception of what eukaryote genomic functions are about. The narrowness is understandable historically: gene functions were the first precisely defined functions that chromosomes were found to serve. Ever since, functions in general were conceived as linked to conserved nucleotides and amino acids.

Functional sequences are sequences whose effects contribute to producing the organism's selected phenotype. A function relates to a performance of the organism based on activities of its component structures. These mostly selected structures and activities are coordinated across all hierarchical levels of biological integration, from molecules and their interactions upward. The term function also designates component structures and their activities at any one particular level to which an observer directs his attention (e.g., macromolecular or physiological function). Structural constituents integrated into functional systems need not all be selected; they can be fixed through other processes as long as they are part of a hierarchically coordinated set of functional effects that is selected.

Neither this nor any other available characterization of functionality enlightens us as to the different ways in which functions are embedded in eukaryote genomes. In examining these ways, particular attention will be given to the following features: genomic functions in eukaryotes are distributed over nucleotide collectivities of various orders of magnitude; the functions specified are linked by their nature to these nucleotide pluralities; and, especially at the higher levels of these pluralities, functions are generated and regulated *epigenetically*. It will be suggested that the functionality or nonfunctionality of a sequence is not definable at a single level of nucleotide plurality. One and the same sequence can be nonfunctional at one level, yet functional at another (and perhaps in more than one way). Given this scale dependence of nucleotide function, it may be difficult to pinpoint truly functionless sequences; namely, sequences that definitely do not carry out even a small fraction of

functional responsibility. The fraction of responsibility can in principle be estimated loosely as sequence fraction of an approximately defined functional aggregate of sequences.

Epigenetic functional effects thrive on sequence aggregates. The term epigenetic will be used in a restricted sense here, and designate heritable changes in chromatin structure unaccompanied by any change in nucleotide sequence (e.g., Hendrich & Willard, 1995; Patterton & Wolffe, 1996; Choo, 2000). 'The links between general epigenetic inheritance mechanisms and chromatin structure and function' (Ferguson-Smith & Surani, 2001) will thus be emphasized. In fact, such epigenetic effects may or may not be functional. The present paper deals with some of those that are.

Epigenetic processes contribute to a 'paradoxical' excess of non-protein-coding over protein-coding sequences. One might distinguish between two forms of the c-value paradox (Zuckerkandl, 1986). Form I relates to minimum haploid genomes in different classes of organisms. To each evolutionary grade will correspond a 'minimum paradox' characterized by a paradox coefficient pc. Pc can be measured as the ratio of haploid genome size over total haploid number of genes. Unfortunately, at present, the haploid number of genes is known with any precision in only a small selection of eukaryotes.

Beyond minimum values for pc, enormous further increases in genome size and in pc recur repeatedly during evolution while increases in numbers of genes remain modest. For each evolutionary grade – in practice, perhaps, for each class of organisms – such increases would come under the heading of c-value paradox II. Polyploidy contributes to genome mass without accentuating the 'paradox'. It does indeed not increase the ratio of haploid genome size over haploid number of genes. Old polyploidies tend, eventually, to evolve into new diploidies, in which cases, barring haploid genome size contractions, it is probable that pc values increase slightly because many genes are presumably turned into pseudogenes. During the process of functional diploidization of polyploidy, the notion of haploidy is blurred.

The minimum c-value paradox (paradox I) is already a full-fledged paradox. The explaining of c-value paradox II seems to be further advanced, thanks to Cavalier-Smith and to Gregory. We shall deal in this paper essentially with c-value paradox I. In accounting for c-value paradox II, some of the functional parameters identified, namely, correlations between total genome size and developmental, metabolic, and even morphological parameters, seem to point in the appropriate directions. Further parameters may not have to be invoked, though mechanisms and causal relationships will have to be further investigated (Cavalier-Smith, 1978, 1982; Cavalier-Smith & Beaton, 1999; Gregory & Hebert, 1999; Gregory, 2001).

In evolution, 'requirements' (say, in regard to length of the DNA fibre) arise only on the basis of preestablished functional relationships. The same requirements would not necessarily be part of a rational and economical plan of a well-informed bioengineer intent on building up a complex living system. To accomplish his task, some evolutionary time would be among his indispensable tools. Because nature proceeds in a quirky fashion, she needs far more time still than would the clever (and long-lived) engineer. Indeed, she tries out everything as though she had no idea of anything. In that she resembles a contemporary computer more than the products of a brain. Upon closer inspection, and contrary to first impressions, the superior powers of nature appear to be brainless. Nature's patent absence of planning – having, in biology, to discover by trial and error even what become her 'methods' – is given a particular relief at the molecular level. Consider, for example, the complexities of 'imprinting' as revealed by Burns et al. (2001). It would be hard to claim that they are all 'necessary' for making an organism function. It has been clear since the seventies at the latest that no new function ever appears in molecular evolution if not through blind tinkering with available structural resources. Under preexisting constraints, a random combinatorial exploration of interactions among components of the system creates new bases for natural selection. This brainless inventiveness may be a prime reason why, in its details, biology is as immensely complicated as it is. In order for tinkering not always to precede selection, it would be necessary for some adaptive features of gene regulation to be transmissible from one generation to the next. In that regard, certain mechanistic possibilities of environmentally directed epigenetic or combined genetic and epigenetic events will, in the future, need to be explored.

One may now have another look at how functions have come to relate to haploid genome sizes. These relationships will be exemplified from Section 4 onward. First, however, we shall discuss two general questions, the relations between function and dispensability of sequences as well as between function and substitution rate.

Dispensability of functional and parafunctional sequences

It is commonly considered that if sequences are found to be dispensable, they are therefore 'junk' (e.g., Kuska, 1998). In fact, one should guard against confusing dispensability with nonfunctionality (Zuckerkandl, 1991). Dispensable genes show that dispensability and function can go hand in hand. Most dispensable proteins probably have been dispensable for a long time, yet been conserved. Dispensable proteins such as human serum albumin (Wilson, Carlson & White, 1977) show all the signs of a long past of honing of their structure and of their sequence, as behooves sophisticated members of the protein community. It is unthinkable that they be structured as they are and that their structures have been conserved as they were if they had not been the targets of natural selection over their long career. They all contain 'essential' evolutionarily conserved amino acid residues. Their sequence variations are in all cases limited to those compatible with a substantial conservation of their tertiary structure (Berezovsky & Trifonov, 2001). In complex organisms, it is unlikely that a significant fraction of proteins have become dispensable suddenly and recently. Besides, even the most dispensable proteins apparently cannot be removed from an organism without at least some residual depressing effect on growth rate, as shown by Figure 1 of Hirsh and Fraser (2001), subsequent to the analyses by Thatcher, Shaw and Dickinson (1998) and Winzeler et al. (1999). The growth-depressing effect may be small, but large enough to account for continued selection of the protein's structure and function over the ages. Alternatively, the effect may be large under recurring environmental conditions differing from those used in reported experiments.

It is true that statistically the most indispensable proteins are the most invariant (Wilson, Carlson & White, 1977; Kimura, 1983; Hirsh & Fraser, 2001). This is perhaps because on average they interact with the largest number of distinct other structures, some of which moreover may be invariant in at least some of their parts (e.g., DNA). Indeed, the evolutionary rate of polypeptide sequences tends to decrease as the number of specific protein/protein contacts (the number of 'interactors') increase (Fraser & Hirsh, in preparation) – the 'Ingram effect' (Zuckerkandl & Pauling, 1965). The most highly dispensable proteins, however, are a mixture in roughly equal parts of variant and invariant proteins (Hirsh & Fraser, 2001). An important fraction of dispensable proteins thus is rather invariant. Nature often resolutely opposes most alterations in a protein and then lets it go altogether. She often acts like an art lover hanging on to a treasure, yet eventually selling it.

The loss of a functional gene may indeed be compatible, in some environments, with the 'reproductive sufficiency' of the species (Zuckerkandl, 1978), namely, the survival of the species is not threatened by the loss. It is all the more to be expected that, under appropriate environmental conditions, reproductive sufficiency will condone also the loss of many noncoding functional DNA sequences. Functional DNA sequences that code neither for protein nor for RNA are often dispensable individually even when they may be indispensable collectively. (Some of the sequence repeats have also individually become indispensable, through special, secondary functional recruitement.) When lost, collectively indispensable repetitive sequences are replaced by different appropriate sequences at the same scale of nucleotide plurality. For example, at a high level of functional nucleotide plurality, a centromere is not lost without being replaced by a neocentromere (Choo, 2000). Individual parts of a mammalian centromere may be lost without the function of the overall sector being compromised, provided that the total sequence length of the overall sector remain above a certain threshold – about 1000 kb (Karpen, Le & Le, 1996). Such a threshold is imprecise and can vary with conditions and with the centromeric function considered. It is collectively that the subsectors of a larger sequence sector provide the substratum for the function(s) (Zuckerkandl, 1986). Therefore, among continuous noncoding sequences acting collectively, no sequences can be individually considered as carriers of the function or be individually considered as nonfunctional.

In essence, individual functioning sequences are sometimes dispensable when they are genes of the kind that multiply by duplication, and are nearly always dispensable as members of retroposition-dependent gene families or as components of higher levels of nucleotide plurality per function. The proper inference to be drawn from acceptable sequence losses regarding sequence function is that no such inference must be drawn.

While collectively certain sequence subsectors are functional, individually they may be termed parafunctional. A parafunctional sequence is one that is not functional by itself, but functions within a collectivity of parafunctional sequences. It must be considered

as functioning as a member sequence, even if overall the collectivity of parafunctional sequences is larger than the minimum size required for the function, for example, when a centromere is longer than needed. In principle, no parafunctional sequence is either more or less functional than any of its sister sequences.

When a noncoding sequence known to be an insert is considered in itself, its functional status is undetermined. It may be functional, parafunctional, or nonfunctional. If treated as junk, parafunctional sequences are entitled to file an antidiscrimination suit.

Functional density and neutral mutation rate

Functional density, a concept that might have more potential than it has been credited with, was defined as the proportion of sites of an informational macromolecule that are engaged in a specific function (Zuckerkandl, 1976, 1986). The counterpart to specific functions of amino acid residues or base pairs are the general functions. Each of the general functions can be represented by a single physical parameter. More than one parameter relates to any residue or base at a given site – for example, in proteins, charge and hydrophobicity. In the case of general functions, substitutions at any one molecular site can routinely be compensated by substitutions at a certain number of other, sometimes distant, sites. General functions involve all sites within the macromolecular entity considered, including the specific-function sites. At sites that fill general functions only, such as a charged site at the surface of a globular protein, namely, when the site is not involved in a specific intermolecular bond, the range of tolerated substituents is relatively wide. On the other hand, specific functions are connected to particular subsets of sites occupied by certain bases or amino acids, with few or no degrees of freedom.

In proteins, indirect effects exercized by amino acid residues at general function sites on specific functions will blur the distinction between general and specific function sites. Thresholds may be set, however, for indirect effects of amino acid substitutions on specific functions, above which a site would be counted among the specific function sites.

Weighted functional density (Zuckerkandl, 1976) is functional density weighted by the mean variability of sites engaged in specific functions. Such weighting is obviously important for establishing a predictable correlation between functional density and evolutionary substitution rate.

In the case of DNA, a distinction between specific and general functions (Zuckerkandl, 1986, Table 1) may be complemented by specifying the order of magnitude of the number of nucleotides involved in a function. When functional density of DNA is low, functionally meaningful nucleotides are spread more thinly over greater sequence lengths and are not individually critical. The rate of accepted mutations is expected to be high and to approach or to have reached the neutral mutation rate. Many sequence patterns, although functionally important, are highly degenerate. One of the weakest such patterns is the nucleosome DNA positioning pattern (Bolshoy et al., 1997). Trifonov and associates have detected statistically several further patterns (e.g., Herzel, Weiss & Trifonov, 1999; Ioshikhes, Trifonov & Zhang, 1999). Their maintenance may express a mixture of weak selection, neutral drift, duplications of nucleotides and of oligo- or polynucleotides, random historical conservation, and coincidence. Whichever process leads to their presence, certain recurring nucleotides appear to be functional and to partake in establishing a value of functional density greater than zero.

Under particular circumstances – DNA strand separation and formation of specific structures in individual strands (Catasti et al., 1999) – functional density measurements would need to include base–base interactions in short repetitive polynucleotide structures – reputed sequence ‘junk’ whose functional density can in fact be high, as the work of Catasti et al., permits one to infer. Short tandem repeats are unlikely to be individually selected. Selection might, rather, intervene by condoning effects of sequence conversion or through the ‘back leap’ phase of a ‘forward creep – back leap’ mode of evolution of a whole zone of repeats (Zuckerkandl, 1975).

In any series of DNA sequences of decreasing functional density, the (locally applicable) neutral mutation rate can be attained well before functional density reaches zero (Figure 1). Indeed, at low functional density, individual nucleotides or amino acids are not expected to be endowed with selection coefficients high enough in absolute value to prevent them under most realistic circumstances from behaving as though they were neutral (Figure 2). Moreover, the neutral mutation rate, along with the mutation rate itself, is expected to vary for various reasons (rates of DNA repair, effects of chromatin structure, etc.) across different organisms as well as across a single eukaryote genome. Thus, in all likelihood, the fact that a region of DNA evolves at what is considered

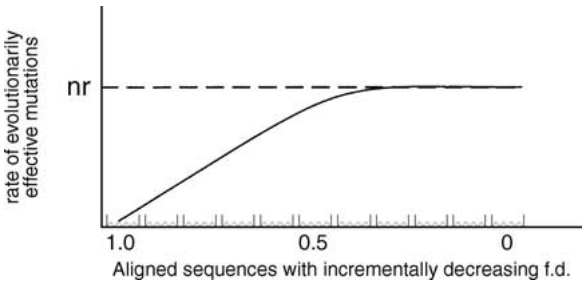

Figure 1. In DNA sequences of decreasing functional density (f.d.), the neutral rate of sequence change is reached before f.d. drops to zero. The abscissa aligns a series of independent sequences of equal length whose f.d. decreases from 1.0 to 0. nr = neutral rate of evolutionarily effective mutations, that is, of alleles represented in a population with sufficient frequency so as to have a significant mean chance of being included in species derived from the species in which they occurred.

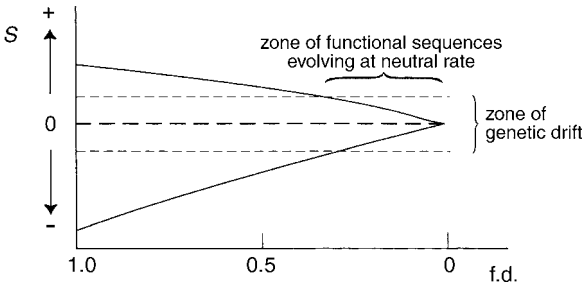

Figure 2. Average positive and negative selection coefficients ***s*** as a function of the functional density (f.d.) of proteins or their coding polynucleotides. Before reaching 0, the selection coefficients become small enough for neutral drift in general to determine the rate of evolutionarily effective mutations (for this term, see legend of Figure 1). The shape of the curve is undetermined – the straight lines are arbitrary. The average ***s*** corresponding to maximal f.d. are represented as larger in absolute value for negative than for positive coefficients, because 'old' proteins are expected to present fewer amino acid substitutions endowed with high positive ***s*** than do the rare proteins that are in the process of perfecting their structure/function relationships on a protein-wide scale (Hartl, Dykhuizen & Dean, 1985; Zuckerkandl, 2001).

a neutral mutation rate tells us nothing about whether that region fills a function or not. It only tells us that the region does not fill a function as a gene. In terms of rate of sequence change, low functional density and nonfunctionality can be expected to give the same result: at low functional density, there can be no selection on a small sequence scale. To be sure, if the term function applies, selection has to intervene at least from time to time, on a larger or, in some cases, on a very large sequence scale – very large when genome size as such may be implicated.

The meaning of the inference regarding equal effects on substitution rate of low and zero functional density deepens as one realizes that there is no good reason to correlate functional density with the 'importance' of a function, even though many biologists tend intuitively to assume such a correlation. Functions of great importance – such as the function of centromeres – may be carried out by sequences of low functional density. In fact, centromeres are the best illustration of the contention that neutral rates of sequence change are reached before function disappears, since 'centromeric repeats comprise the most rapidly evolving DNA sequences in eukaryote genomes' (Henikoff et al., 2001). Those who throw these sequences into the garbage are not just throwing away DNA; they are throwing away most of the eukaryotes.

Let us first comment on sequences of high functional density, which are also the sequences that in DNA are selectable over the shortest sequence length.

Polynucleotide function on a nucleotide per nucleotide scale

As soon as a correlation between function and individual nucleotides does not obtain, investigators often look askance at function as though on a homeless drifter; on a genetic drifter, as it were. Correlations between functions and individual nucleotides are classically considered to be limited to the small scale of coding sequences, their associated promoters and enhancers, and a few other types of relatively short sequences such as insulators (Jackson, 1997; Sigrist & Pirrotta, 1997; Wolffe, 2000; Bell, West & Felsenfeld, 2001). In addition, short non-protein-coding conserved sequences permeate eukaryote genomes. While part of these conserved sequences may function in *cis*, others may function via transcripts (see note added in proof). The connection between conservation and function seems much tighter than the erroneously touted connection between nonconservation and nonfunction. In the neighborhood of genes, 20–30% of the noncoding sequences are conserved in distantly related *Drosophila* species (Bergman & Kreitman, 2001). An analysis covering the genomes in a relatively comprehensive way has now shown that, overall, roughly one third of the human and mouse intergenic nonrepetitive sequences (of which about 19% are transcribed though not translated) are composed of constrained nucleotide strings of a mean length of about 134 nucleotides (Shabalina et al., 2001). The fraction of constrained nucleotides decreases as the length of the intergenic regions increases.

Nevertheless, the number of constrained nucleotides goes up as the nonrepetitive non-protein-coding genome expands. Therefore, even in large genomes, deletions in non-protein-coding sequences must be relatively small if they are not frequently to include constrained nucleotides. At this point, if it were sentient, the c-value paradox would begin feeling a little uncomfortable. But its travails are not over.

There is, indeed, a quantitatively important class of high-functional-density DNA sequences that were left out in the analysis quoted, and are thought to be genomic parasites having multiplied exclusively for their own benefit, not for the organism's. They are the SINEs and LINEs (Schmid & Maraia, 1992; Jurka, 1995; Smit et al., 1995; Schmid, 1996; Furano, 2000), namely, middle-repetitive sequences such as the shorter retropositionally dependent Alu and the longer retropositionally independent L1 'non-LTR' retrotransposons (e.g., Eickbush, 1992). The retrotransposition of SINEs depends on the ORF-2 gene of LINEs, which encodes a transposase that functions for both the LINE from which it originated and for SINEs (Boeke, 1997; Weiner, 2000).

SINEs carry conserved internal RNA polymerase III promoters, though many SINEs are no longer or have never been transcribable. Originally, it was easy to deny to Alu sequences the exercize of any organismal function in *trans* (let alone in *cis*), because their polymerase III transcripts had not been detected *in vivo* (cf. Schmid & Shen, 1985). Also, the number of SINEs in some genomes just seemed too large for any functional rationalization. (Yet, one might remember how many spermatozoa are available to fertilize sometimes a single egg. The excess does not render spermatozoa nonfunctional.)

SINE and LINE genes differ from most other genes by their capacity for selfpropagation (in the case of SINEs, with the assistance of a helper LINE) and for dispersal over the genome. For a long time this difference served as a pretext for dismissing them as genes except in an exclusively selfish sense. However, many of the SINEs and LINEs that have lost the ability to be retrotransposed continue to be transcribed as functional genes would. Indeed, in a functionalist vein, SINEs were considered to be 'cheap genes' (Zuckerkandl, Latter & Jurka, 1989) – genes that come cheap from the point of view of selection. It appeared subsequently that they did not come as cheap as had been assumed: a number of nucleotides in Alu sequences are conserved during evolution (Britten, 1994, 1995).

The action of natural selection on a few retrotransposable 'source' genes had already been inferred. But after the work of Britten, natural selection appeared to act on a very large number of Alu elements. An overwhelming majority of these elements are not retrotransposable and therefore could have no selfish reason for undergoing selection. Might mutational or recombinational cold spots, rather than selection, be responsible for the conservation? These two alternatives would require that by extraordinary coincidence either 'freezing' process has led to the conservation precisely of nucleotides known to be functional transcriptionally. The conserved nucleotides, indeed, provide for both transcriptional repressibility and competence. Regarding repressibility, binding sites for a 'strong' nucleosome are apparently conserved (Englander, Wolffe & Howard, 1993); and as to transcriptional competence, so are sites of the polymerase III promoter (Kariya et al., 1987; Britten, 1994). Mechanisms for transcriptional activation and repression that are maintained by selection in non-retrotransposable SINEs must have function(s) that are other than selfish. There hardly seem to be practicable routes of escape from selection at the service of the organism.

Unquestionably, SINEs and LINEs do have an aspect of parasites (Doolittle & Sapienza, 1980; Orgel & Crick, 1980), but their parasitism is ambiguous. Unambiguous parasites are those, one might say, for which horizontal transmission is a habit. There is no evidence for horizontal transmission of SINEs and LINEs between individuals or species (Malik, Burke & Eickbush, 1999). Evolutionarily, the retroviruses, which are clear-cut parasites, apparently descended from molecules that were not clear-cut parasites and that resembled LINEs: outright parasitism very likely came second (Doolittle et al., 1989; Xiong & Eickbush, 1990; Eickbush, 1992). Not only retroviruses, also LTR retrotransposons are younger than non-LTR retrotransposons and seem to be derived from the latter (Malik & Eickbush, 2001). Unambiguous parasitism does not seem to appear at the root of the tree. No reason for surprise, if it is realized that parasite-related entities and organismic function do not exclude one another. They certainly are strange bedfellows; but bedfellows they often are.

SINEs may well have simply exploited and modified an ancient predisposition of tRNA genes to disperse over genomes as retrotransposons (cf. Maraia & Sorrowa, 1995). Nor should such dispersion disqualify SINEs from being treated as true genes.

It is sometimes held that SINEs and LINEs occupy in genomes as much space as they do simply because it would be too difficult and too costly for the organism to prevent the selfish and parasitic retrotransposons from spreading. This argument appears suspect, in view of the fact that, in somatic tissue, the vast majority of transcriptionally competent Alu elements can be and usually is masked from the polymerase III transcriptional machinery (Russanova, Driscoll & Howard, 1995). Then why does the organism endorse their enhanced activity in the male – though not in the female – germline (Miller, 2000)? Also, certain organisms, such as filamentous fungi, have been able to keep the number of their repetitive sequences at very low levels through repeat-induced point mutation (RIP) and methylation induced premeiotically (MIP) (cf. Henikoff & Matzke, 1997) – devices that under some form might have been adapted to other genomes. Presumably, multicellular eukaryotes refrain from resorting to general and permanent suppressing measures because such measures would lead to throwing the baby out with the bathwater.

Evolution being recognized as the supreme opportunist, it is not disputed that retrotransposons, here and there, play in gene regulation some secondarily acquired roles in *cis* (Brini, Lee & Kinet, 1993; Britten, 1996a, b, 1997)– roles of the genetic, not the epigenetic kind. Published and unpublished work from Carl Schmid's laboratory now supports the presence in SINEs and LINEs of pervasive and ancient phenotypic functions in *trans*.

LINEs are functional at the very least because they ensure the genomic distribution of SINEs. To be sure, this particular function of LINEs (unlikely to remain the only one in *trans*) is predicated upon SINEs being generally functional in the first place. One additional role for LINEs has already been claimed, namely, in cell proliferation (Kuo et al., 1998). Among recent observations on LINEs, one may highlight the following. L1 sequences or other LINEs are ubiquitous in eukaryotes[1] and are about as old as the eukaryotes. Non-LTR elements are in general as old (Malik, Burke & Eickbush, 1999). In multicellular eukaryotes including plants (Schmidt, 1999), LINEs can be regulated so as to be transcribed in certain cell types and severely repressed in most tissues (Tchenio, Casella & Heidmann, 2000). In fact, 'L1 expression is controlled by a tightly regulated temporal and spatial program of events during development' (Trelogan & Martin, 1995). We thus have antiquity, ubiquity, and seem to have regulation.

Massive numbers of LINE or LINE-related sequences are known to exercize functions in *cis*, namely, structural functions with respect to chromatin structure. A certain linear frequency of old LINE 1 (L1) sequences would appear to be causally linked to the facultative heterochromatinization of the human X chromosome (Bailey et al., 2000; Lyon, 2000). In other organism, too, retrotransposons comparable to LINEs collectively fill important functions in relation to chromatin structure – namely, the HetA and TART repeats in *Drosophila* telomeres (Pardue et al., 1997).

Though partly still circumstantial, a case for functions in *trans* of *SINEs* has now become strengthened by evidence from the laboratory of Carl Schmid. Schmid and his associates (Chu et al., 1998; Kimura, Choudary & Schmid, 1999) presented data in support of the view that SINE RNAs serve a role in cell stress response, a role that predates the divergence of insects and mammals. SINEs are thus considered to represent a class of cell stress genes, and very old ones at that. Under cellular insults (viral infection, heat shock, etc.), the abundance of full-length Alu (flAlu) RNA increases by as much as 50-fold. The overexpressed flAlu RNA stimulates protein synthesis and inhibits the activity of a general repressor of protein synthesis, PKR, which is an eIF2 (translation initiation factor 2) protein kinase (Chu et al., 1998). flAlu RNA binds to PKR with high affinity (Chu et al., 1998; Schmid, 1998). Regarding evolutionary origins, a homology relationship has been reported between PKR and tRNA synthetases (see Schmid, 1998). There is an apparent 'very ancient association between a primordial PKR's RNA-binding properties and the very deepest evolutionary roots of SINEs within the tRNA superfamily.' Early during the evolution of multicellular eukaryotes, tRNA-derived SINE transcripts would have specialized as interacting partners of PKR, and, along a certain line of mammalian descent, this partnership would much later (before the common ancestor of rodents and primates, Jurka & Zuckerkandl, 1991; Quentin, 1994) have been taken over by Alu sequence transcripts. To be able to play that role, Alu sequences are not only sufficiently related in secondary and tertiary structure to tRNA-derived SINEs thanks to their cloverleaf conformation (Okada, 1990; Maraia & Sarrowa, 1995), the two types of SINEs

[1] In a group of South American rodents, functioning L1 sequences seem to be extinct (Casavant et al., 2000). One might conjecture that, in these rodents, the role of L1 has been taken over by other LINEs; or that the long-term welfare of these species may be in question.

are perhaps homologous contrary to current belief[2]. A long evolutionary continuity of a particular regulatory relationship would thus have been maintained.

This view, based on a considerable body of experimental work, is bolstered by observations on the protist *Tetrahymena* in which heat shock induces the rapid accumulation of a 7SL-related, polymerase III-transcribed small RNA, an RNA reminiscent of Alu sequences and required for the establishment of thermal tolerance (see Chu et al., 1998).

The transcriptional activation of SINEs as a stress response has also been observed in tRNA-derived SINEs, be it in mice or in silkworm (Kimura, Choudary & Schmid, 1999; Li et al., 1999; Kimura et al. 2001). This observation fits in with the structural commonalities between tRNA-derived and 7SL-derived SINEs. Moreover, 'all known active plant retrotransposons are largely quiescent during development but activated by stresses, including wounding, pathogenic attack, and cell culture' (Wessler, 1996).

It is up to those who remain committed to the selfish gene paradigm for SINEs and LINEs to explain how a purely parasitic type of sequence can put on so successful an act in counterfeiting a type of molecule of general impact on translational control, a type of molecule that, under different guises, has been on the job for a remarkably long time.

Beside translational control, Alu sequences, under stress conditions and in *trans*, may also exercize transcriptional control, since PKR reportedly phosphorylates the transcriptional regulatory factor I-κB, an inhibitor of members of the NF-κB/c-*rel* family. These members are active in growth regulation, differentiation, and other fundamental processes (see Clemens, 1996).

In addition, as will be discussed presently, it is to be expected that Alu sequences routinely have inhibitory effects in *cis* on 'ordinary' genes. Here again, the structural analogy with tRNA-related SINEs is of potential functional import: both types of SINEs appear to promote methylation in neighboring genes (Hasse & Schulz, 1994).

Overall, considering the available set of observations, however incomplete it be, we have once more, in the case of SINEs, antiquity, ubiquity, and regulation – in *trans* and in *cis*. On both counts, *trans* and *cis*, more functional connections remain to be discovered. Some such connections in *cis* will now be put in focus, because they concern the contribution of presumed junk DNA to epigenetic control.

SINEs as a link between the genetic and epigenetic regulation of genes

The introduction of epigenetic control systems into the living world is extremely ancient, even though the 'sectorial' mechanisms to be mentioned may be limited to the eukaryotes. A number of genes in *Escherichia coli*, however, are subject to epigenetic controls, whose mechanisms include methylation (in this case of adenosine rather than of cytosine) (Henderson, Owen & Nataro, 1999).

Many Alu sequences are sites of epigenetic processes. In somatic tissues and in oocytes, Alu DNA is heavily methylated. The transcription of methylated Alu sequences is inhibited by a repressor (Liu et al., 1994). On the other hand, certain Alu subgroups are almost completely unmethylated in sperm (Liu et al., 1994; Schmid, 1996). 'The demethylation of a major Alu subset in sperm almost certainly derepresses their transcription following fertilization' (Schmid, 1998), presumably establishing an allelic difference in transcriptional behavior during embryogenesis, a difference characteristic of 'imprinting.' There would be little promise in trying to construe imprinting as a feature of the selfish behavior of parasitic sequences. Imprinting is distinctively phenotypic in its thrust. In general, the impact of imprinting is on the functioning of organismal genes, particularly during embryonic development. 'Imprinting is unlikely to arise unless there is some selective advantage to being imprinted' (Spencer, Clark & Feldman, 1999).

[2]It does not seem unlikely that the tRNA-like sequence in 7SL RNA was in fact derived from a tRNA. Similarity in secondary and (as presumed here) tertiary structure (Maraia & Sarrowa, 1995), though not demonstrative, still represents a point in favor of homology. With the help of other structural and functional features of the 7SL RNA molecule, the tRNA moiety of 7SL may have been able to alter its original sequence while conserving much of its structure. Subsequently, sequence readjustments (Okada, 1990; Quentin, 1994) would have been necessary in order for a primate Alu or a rodent B1 sequence precursor to be weaned from the structural and functional support conjectured to have been dispensed by partner sequences within 7SL. For example, in 7SL RNA, relative to lysine tRNA, the B-box of the polymerase III promoter was destroyed through a large insertion (cf. Okada, 1990) and had to be reestablished in the ancestor of the human Alu left monomer through a reinsertion of the trinucleotide GAG (actually the tetranucleotide GAGA) (Okada, 1990). The sojourn of the SINE precursor sequence within 7SL RNA would have helped efface convincing evidence of the extant sequence homology between tRNA and future Alu and B1 elements.

Beyond a probable source of imprinting, SINEs are also a source of position effect variegation (PEV). The underlying mechanism and its implications can be described in several steps.

Repeat sequences in euchromatin can lead to heterochromatinization

Tandem repeats in euchromatin – especially in inverted configuration – lead to local heterochromatinization. This has been demonstrated at least for fairly long repeat units (about 10 kb) of which three to four in a row suffice for forming a heterochromatin-like structure (Dorer & Henikoff, 1994, 1997; Henikoff, 1996). It is well known that very short repeats as found in satellite sequences, when numerous enough, can form heterochromatin through interactions with particular proteins (e.g., Zuckerkandl & Hennig, 1995; Eissenberg & Hilliker, 2000). For example, heterochromatin protein 1 (HP1), present notably in constitutive heterochromatin, also binds to repetitive retroposon arrays (Fanti et al., 1998) – a telltale connection.

Heterochromatinization can lead to position effect variegation

Constitutive heterochromatin induces position effect variegation (PEV) in euchromatic genes brought into contact with it by transposition (Lewis, 1950; Spofford, 1976; Weiler & Wakimoto, 1995). This ability of constitutive heterochromatin can be extended to tandem repeats located within regions defined as euchromatin (Henikoff, 1996). In this version of PEV, local 'heterochromatinization' can occur in the absence of any transposition. For example, 15–20 inserted copies of a transgene caused cellular mosaicism in the methylation and expression of hemizygous loci in the mouse (McGowan et al., 1989). Robertson et al. (1995) reported on a transgene construct whose expression is variegated when it is present in multiple copies, and Wallrath (2000) found that variegating transgenes in *Drosophila* were all located adjacent to repeat elements. By further analyses of the expression of transgenes in *Drosophila*, strong support has been given to the view that middle repetitive sequences, when grouped, can form a heterochromatic region in the presence of appropriate factors, notably heterochromatin protein HP1, and lead to mosaic gene expression (Seum et al., 2001). In the presence of arrays of human Alu elements, the Alu-binding factor YY1 is among the proteins whose local distribution might intervene in determining the occurrence and in modulating the stability of a heterochromatic structure (Humphrey, Englander & Howard, 1996); and therefore in determining the occurrence and extent of ensuing position effect variegation.

Sequence repeats in euchromatin pair with constitutive heterochromatin through DNA looping

What triggers the relative stabilization of a heterochromatin-like structure in sequence repeats located well within a heretofore purely euchromatic domain? Thanks to DNA looping, repeat sequences in euchromatin join up and pair with constitutive heterochromatin, even at considerable distances (Dorer & Henikoff, 1997). Thus, a process, heterochromatinization, that had been abnormal in the case of transposition becomes potentially normal through looping. Given the large number of multiple sequence repeats present in mammalian genomes, repeats that often are not far apart, 'normal' PEV can be considered as potentially frequent.

Cis-*acting SINEs offer an additional pathway for transcriptional regulation*

Within previously euchromatic sectors of DNA that have since been the targets of neighboring insertions of middle-repetitive sequences, repeat-induced PEV might thus vary from slight to strong, in terms of the percentage of cells in which the transcription at least of the gene closest to the repeat area is repressed. This process could amount to an 'analog' (or 'rheostatic,' Fiering, Whitelaw & Martin, 2000) method of transcriptional regulation during development at the level of the overall performance of a cell population or a tissue. In the individual cell, on the other hand, PEV inhibition or absence of inhibition of a gene appear to be digital and binary (Spofford, 1976; McGowan et al., 1989; Singh, 1994; Felsenfeld et al., 1996; Ma et al., 1996; Henderson, Owen & Nataro, 1999; Lloyd, Sinclair & Grigliatti, 1999; Rossi et al., 2000; Sutherland et al., 2000).

The effectiveness of repeats in bringing about PEV may be opposed by sequence elements that protect the activity of a gene, such as enhancers (Walters et al., 1996; Francastel et al., 1999), as well as by competing positive transcription factors when they are present early enough in development (Ahmad & Henikoff, 2001). If PEV wins out, mature tissue is composed of mosaic cell clones. The proportion of cells sporting the gene in its repressed state will depend, among other parameters, on suppressors and enhancers of

variegation which may vary in amount with cell type, developmental stage, and body location. The potential thus exists for epigenetic gene regulation to be adjustable as a function of a cell population's state of differentiation and position in local gradients. The imprinted and variegated H19 gene offers an illustration, along with the reciprocally imprinted Igf2 gene. During development of the mouse, these genes are regulated through the intervention of different enhancers in different cell types (Ainscough, Dandolo & Surani, 2000). This implies the use in different cell types of different factors or assortments of factors.

A gene's distance from stable heterochromatin affects its degree of variegation (Dorer & Henikoff, 1997; Lloyd, Sinclair & Grigliatti, 1999). Therefore, the amount of intervening 'junk DNA' has a regulatory function reflected in the level of expression of a gene in a tissue as a whole. Increases in genome size through insertions might well lead to an attenuation of PEV, namely, when the inserts occur between repeat sequences and thus increase their mutual distances – except perhaps if the inserts contain themselves a similar repeat. This would be an example of a possible direct effect of changes in genome size on gene regulation.

Hypothetically, a regulatory role in *cis* can thus be attributed in a significant number of cases to what is in the mind of many the most typical DNA 'junk', namely to old SINEs and LINEs that are no longer transcriptionally active and to 'dead-on-arrival' truncated SINEs or LINEs, if they have conserved sufficient sequence similarity among themselves. That role would fade out progressively over evolutionary time as the similarity and thus the ability of the repeat sequences to be heterochromatinized is gradually effaced by accepted mutations – unless the cluster of repeats is refreshed by new arrivals or their decay is limited by selection.

Because of the great frequency in genomes of repetitive sequences, there may be many more variegating genes than has hitherto been recognized. It would no longer be appropriate to write 'PEV is a result of aberration and is not part of a cell's normal physiology' (Singh, 1994), and the seemingly strange fact (Singh, 1994) that heterochromatin protein 1, HP1, also binds to a number of euchromatic sites (Fanti et al., 1998) would no longer be unexpected.

Thus, an additional avenue for the evolution of gene regulation is provided through the introduction or loss of repetitive sequences (loss by instantaneous deletion or slow mutational drift). In this way, many more sequences may well be involved in gene regulation than are commonly thought to be, and repeats of 'functionless' sequences, in addition to being pathogenically disruptive in a number of instances (see Schmid, 1998), may indeed have functions of an adaptive kind.

There is much evidence to show that *the eukaryote genome is an epigenetic machine*, besides being a genetic one. One important argument in support of the proposition that most DNA is junk has been the presence in genomes of very large amounts of 'meaningless' repetitive sequences. In fact, not only tandem repeats of simple sequences as in heterochromatin or (a matter not explored here) in mini- and microsatellites, also dispersed middle-repetitive sequences can support epigenetically controlled functions.

Functions of nucleotide pluralities of the order of 100 kb

Some functions are linked to continuous sectors of DNA on a scale much larger than that of average individual genes. Different functions can be correlated with nucleotide collectivities of different sizes. The usual nucleotide scale involved in genes and in their local ('punctate'[3] as contrasted with 'sectorial', Zuckerkandl, 1997) transcriptional control is 1–10 kb, with the majority of sequences non-protein-coding, namely, primarily introns. At the next magnitude level of 100 kb, single genes, split apart by very long introns, may occasionally be encountered (Deutsch & Long, 1999). Conserved noncoding sequences of similar sizes are also found, as around the genes for the mammalian T cell receptor loci (Koop & Hood, 1994; Koop, 1995).

Certain cases of PEV involve nucleotide counts of the same order, and so does imprinting (Ainscough et al., 1998). The same nucleotide scale is relevant to a level of gene control that I had addressed in the mid seventees under the vocable of eurygenic ('wide') control through mass binding of factors to chromatin – later referred to as sectorial binding – as distinct from stenogenic ('narrow'), more local control through punctate binding. Since then, gene sequestration through sectorial factor-binding has come to be

[3] Regulation can be called punctate even when the distance between a promoter and an enhancer is of the order of 100 kb, because interacting promoters and enhancers still represent only one 'point' in chromatin space.

considered an essential component of eukaryote development (cf. Zuckerkandl, 1997, 1999; see also Caplan & Ordahl's, 1978, 'irreversible repression' during development). Such sequestration may correlate with a specific differential distribution of sectorially repressed genes during interphase over different nuclear compartments, according to developmental phase and cell type (Csink & Henikoff, 1998a; Brown, 1999; Marshall & Sedat, 1999). The use in gene regulation of considerable lengths of non-protein-coding DNA is in part understood, in part tentatively understood, and raises additional questions.

Position effect variegation, imprinting, and cell determination are aspects of a unitary epigenetic process resulting in gene sequestration

The most notorious example of gene sequestration during development as a function of body region and time is presented by the 'superrepressible' (sectorially repressible) Hox genes (Lewis, 1978; McGinnis et al., 1984; Scott & Weiner, 1984; Gaunt & Singh, 1990; Duboule, 1992). Sectorial repression occurs through interactions of chromatin with a set of sectorially binding factors, the polycomb group (Pc-G) factors. (The name designates *Drosophila* factors, but there are homologs throughout the eukaryotes). The same sector of chromatin binds another set of factors, the trithorax-group (trx-G) factors, thanks to which, in different body locations, different subsectors of the overall superrepressible sector remain, on the contrary, poised for activation. While sectorial repressibility is found in gene clusters, it may well also apply to isolated regulatory genes. Note that the molecular mechanism of sectorial repression in terminal differentiation genes (e.g., the β-globin genes) differs from the mechanism in key developmental genes (Li, Liu & Liang, 2001).

Superrepression may be considered the corner stone of cell determination (Zuckerkandl, 1999), which consists in laying down in chromatin the high-order structural foundation for cell types. Cell determination is the framework for differentiation. During differentiation, particular programs of gene action are activated in preference over others. Determination appears as a double-facetted process, with a genetic and an epigenetic side. Neither of these facets could be present without the other; nor, so it would seem, could the epigenetic side of cell determination exist without the dynamic structural provisions that lead to PEV.

There is a strong reason why PEV is a general biological 'institution' in eukaryotes: one and the same developmental gene, if superrepressible, is either sectorially repressed or sectorially poised for activation. Repressing and repression-preventing factor complexes preempt each other's effectiveness at certain moments of development, in accordance with Felsenfeld's (1992) 'preemptive competition' model. In general, and within factor concentration limits (Zink & Paro, 1995), who gets there first at the right time takes possession of the chromatin sector, not only in one particular cell, but in this cell's progeny (e.g., Poux, McCabe & Pirrotta, 2001). Variegation is thus an integral part of gene sequestration.

Imprinting, in turn, is essentially the expression of differential sectorial repression of a given gene in the male and female germ lines, with one of the germ lines sequestering the gene through sectorial repression, while the other germ line makes it available to punctate regulatory factors, so that the diploid cell resulting from fertilization is functionally haploid for the gene. Sapienza (1995) already pointed to an intimate relation between imprinting, heterochromatinization (a process connected with sectorial repression), and PEV.

Expression can be complementary to repression. Spatiotemporally bounded conditions for sectorial repression shared by neighboring genes might occasion corresponding similarities in the genes' activity patterns.

Gene sequestration and introns

In sectorial repression, coding sequences and their promoters are not expected to be able to sequester themselves. They need noncoding sequences to do the job (e.g., Zuckerkandl, 1974, 1999; Orlando & Paro, 1993; Pirrotta, 1997). Introns and flanking sequences could participate effectively in compensating for a presumably mediocre ability of exons to form stable higher-order structures on their own. This relative lack of structural competence on the part of coding sequences would be attributable to their commitment to other priorities. Therefore, the proposal was made that a general function of introns consists in their potential to help stabilize those local high-order structures of chromatin in which genes, at times, have to be sequestered (Zuckerkandl, 1981; Beckman, Brendel & Trifonov, 1986; Beckman & Trifonov, 1991).

Consequently, in the course of evolution toward more complex organisms in which cell-type-specific 'superrepression' plays an increasing role, one would anticipate introns to be inserted much more frequently than removed. This is the case (Cho & Doolittle, 1997; Rzhetsky et al., 1997; also Stoltzfus et al., 1994). Actually, in this context, an epidemic of intron insertions is not expected to have broken out before high-order chromatin structures were used for stable transcriptional repression.

The oldest introns nevertheless probably originated in another circumstance, the formation of complex proteins by the fusion of smaller protein modules (de Souza et al., 1998). The controversy around the introns-early/introns-late theories has still not been put completely to rest (Wolf et al., 2000). Yet, it is generally accepted that a large fraction of introns have been secondarily inserted into genes.

A stabilizing and sequestering role of introns would predict that the larger the sum of a gene's exons, the larger will be on average the summed size of the introns. On the other hand, small genes would in general have no introns. On the basis of a limited (and in part redundant) set of data, indications in favor of both points have been obtained (Naora & Deacon, 1982a). As likewise expected, the length of intergenic sequences in gene clusters also seems to increase as the coding sequences become longer (Naora & Deacon, 1982b). For a total coding sequence length of about 0.6 kbp, a gene was (tentatively) calculated to come with a territory of a 24-fold larger amount of noncoding sequences.

Likewise, a foreseeable correlation between the size of introns and the size of either preceding or following exons has been found to be significant, though weak (Smith, 1988). Total intervening DNA sizes per gene and mRNA sizes are also positively correlated (Smith, 1988). Extreme exon sizes seem to be eliminated by purifying selection.

In harmony with a structure function of introns, Beckman and Trifonov (1991) found that the distances among the beginnings of introns and those among the ends of introns are such as to ensure a correspondence with the length of nucleosomal repeats. Genes deprived of introns might no longer be able to form well-ordered nucleosomal arrays (Liu et al., 1995) – a circumstance that would interfere with heterochromatoid structures in superrepressed genes.

Denisov, Shpigelman, & Trifonov (1997) observed a tendency for the strongest nucleosomes to form most often in the introns, and Levitzky et al. (2001a) showed that introns exhibit higher 'nucleosome formation potentials' than exons: efficient nucleosome positioning sites occur preferentially in introns. A fraction of exons has an even lower nucleosome-binding capacity than random sequences do, an observation that might have been predicted. These findings again support the conjectured chromatin-organizing role of introns and the, on average, definitely diminished ability in this respect of exons.

'Open' housekeeping genes characterized by CpG islands in GC-rich isochores would not be expected to use introns to the same extent for ensuring the stability and regularity of a higher-order chromatin structure. It has in fact been found in warm-blooded vertebrates that genes in the most GC-rich isochores, known (cf. Bernardi, 1995) to be housekeeping-gene-rich, contain on average the shortest sums per gene of intervening sequences (by a factor of 3) as well as the shortest intergenic sequences (Duret, Mouchiroud & Gautier, 1995). Equally significantly, the nucleosome formation potential of promoters differs markedly according to whether the genes are 'housekeeping', 'widely expressed', or tissue-specific (Levitzky et al., 2001b). The nucleosome formation potential of promoters is by far the highest in tissue-specific genes, many of which must be sectorially repressible (Zuckerkandl, 1999) as are many developmental regulatory genes, and should indeed have the strongest requirement for introns. The potential in question is the smallest in housekeeping genes, whose promoters actually are found to exclude nucleosomes to a far greater extent than random sequences do. It will be of interest to see whether the average intron number and size per gene is decreased in housekeeping and 'widely expressed' genes as expected.

In summary, as coding sequences structure their residence in chromatin, they not only fragment into exons and introns for an apparent 'reason'; for the same reason, they moreover seem on average to require more linear space on the noncoding flanks of the coding regions. These are roles for 'junk DNA' that, for a large part, are yet to be explored. The contribution of introns to the stability of high-order chromatin structures probably represents only one of the determinants of intron and exon size and distribution. For example, intron size varies in step with genome-wide insertion and deletion equilibria (Moriyama, Petrov & Hartl, 1998). Intron sizes are thus changed mutationally before there can be any question regarding roles that such size changes may play.

Even nonalignable sequences may share similar structural and functional properties

How about conserved and nonconserved nucleotide runs in sectorially repressed genes? In *Drosophila*, the *engrailed* (*en*) gene, like the Hox genes, is superrepressible and is the target of Pc-G and trx-G factors. In the *en* locus, 30% of the noncoding sequences are conserved between *Drosophila melanogaster* and *Drosophila virilis* whose common ancestor dates back to 60 million years ago (Kassis, Wong & O'Farrell, 1985; Kassis et al., 1989). The conserved sequences consist in short runs of up to, say, 40 bp and sometimes more. Seventy percent of the locus is variable.

PREs, namely, polycomb group protein-responsive elements (e.g., Orlando & Paro, 1993; Paro & Harte, 1996), provide nucleation centers for the high-order structure characteristic of the repressed state of Pc-G-controlled superrepressible genes. The factor complexes formed around and between the PREs are made of multiple components, several of which apparently participate in a large proportion of PREs. PREs may not be linearly homologous among themselves, but present a modular composition of sequence motifs (Pirrotta, 1997; Horard et al., 2000). Some of the motifs may be present in nearly all PREs (Strutt & Paro, 1997; Mihaly, Mishra & Karch, 1998). PREs induce position effect variegation (Chan, Rastelli & Pirottta, 1994; Zink & Paro, 1995; Pirrotta, 1997; Horard et al., 2000), probably because the choice between either forming or not forming the sectorially repressive structure frequently depends on moderate changes in factor availability and state, conditions that may fluctuate from cell to cell.

Once the anchoring of a Pc-G complex has been accomplished through a member of the complex – namely, Polycomb protein in experiments of Müller (1995) – other members of the complex will accrete to that member. When, subsequently, the anchoring molecule is removed, the complex not only remains in place, but is stable over many cell generations, provided that the flanking sequences contain a PRE (Müller, 1995). PREs are, however, necessary only for maintenance throughout development and for heritable transmission of the repressive factor complexes (Kassis, Wong & O'Farrell, 1989), not for their formation and their sequestering action. The formation of the repressive structure would seem compatible with a variable noncoding sequence neighborhood, as long as a 'seed' Pc-G factor is tethered to the DNA (Müller, 1995). The anchoring to DNA is required as a local structural trigger.

It is doubtful that an alignment of Pc-G-controlled noncoding sequences from the same species would display homology overall. Yet, not only the conserved, but also the variable sequences, which are in the majority, seem to collaborate in the same structural and functional effect of superrepression. Similar results obtained with dissimilar sequences could be explained by cooperative effects of factors that, among themselves, contain the largest part of the information required for forming the structure. The greater part of structural specificity of multimeric protein/DNA complexes would reside in the proteins.

In the case of nucleotide pluralities per functional sequence of the order of 100 kb, the contribution of relatively high-functional density DNA sequences (PREs etc.) is still conspicuous. Yet, as will be suggested, at even higher levels of nucleotide plurality per function one can observe again the primacy of mutually interacting protein factors over DNA sequence in determining chromatin structure. At these levels, the contributions of short, higher-functional-density DNA sequences will be less prominent. Apparently, the phenomenon of a shift of specificity from compounded (DNA sequence specificity plus factor sequence specificity) to a regimen of decidedly predominant specificity of factor sequences becomes more conspicuous with rising levels of nucleotide plurality per function.

The binding of common groups of factors over chromatin sectors lacking extensive sequence similarity may be widespread in genomes of multicellular organisms. It looks as though structural and functional similarity were inducible at times when sequence similarity is lacking. Implicit in this view is that nonalignable sequences may share similar structural and functional properties.

Conserved segments of noncoding sequences certainly can fill *cis*-regulatory roles. However, these segments are likely to relate only to punctate gene regulation or to levels of sectorial functional regulation at which the contribution of high-functional density sequence motifs is still marked. Punctate gene regulation goes hand in hand with conservation of non-protein-coding sequences. Sectorial regulation, which provides an additional and more nucleotide-consuming aspect of gene regulation or regulation of other types of function is not dependent upon sequence conservation to a comparable extent.

In sectorial gene regulation, as well as in functions using even higher orders of nucleotide pluralities, non-alignable noncoding sequences can be *cis*-regulatory (or, in a more general sense, *cis*-functioning), just as alignable sequences are. This is why it is so difficult to interpret in functional terms findings such as those of Shabalina and Kondrachov (1999). The authors compared sequences of two species of the flatworm *Caenorhabditis* believed to be distant enough for most sequence similarity to have been lost in unconstrained sequences. Among introns and intergenic sequences, the authors observed a mixture of rather strongly and moderately constrained sequences. They recorded 17–18% on average of constrained nucleotides in intergenic sequences and in introns, and believe the remainder of the sequences to be 'truly functionless.' This judgment is standard and seems too facile. It is based in particular on the occurrence of many accepted insertions and deletions. However, insertions and deletions may not be disruptive of functions linked to higher order pluralities of nucleotides. In some cases, on the contrary, indels, as we shall discuss, may contribute to getting a longer sequence 'elected' for a function that it had not previously carried out.

Is distance between enhancers and promoters a functional character?

As one considers sectors of DNA in which nucleotide sequence becomes increasingly unimportant, the last function left could be distance. The function of distance in genomes requires further investigation. Some of the genomic distances involving the 10–100 kb areas of function may be considered briefly.

An important aspect of the complexity of gene regulation and gene interaction in higher organisms is provided by the existence of multiple enhancers per gene. The spread and multiplication of enhancers over relatively considerable distances from their cognate coding sequences may be commonplace at least in mammals. For the imprinted genes Igf2 and H19, which are 70 kb apart, two enhancers were known downstream of H19. A search for more enhancers came up with a total of 10 over 40 kb (Ishihara et al., 2000). There may be more (Ainscough et al., 2000). In the case of the *Bmp5* gene (for bone morphogenetic protein 5), enhancers were found to be located up to 270 kb downstream of a gene whose expression they partly control (DiLeone et al., 2000). Function may be compatible with considerable variation in these distances. On the other hand, the distances might not be indifferent from a regulatory point of view (Casares et al., 1997).

At the low end of these distances, the successive action of enhancers in the transcriptional regulation of a gene no doubt necessitates a minimum tether lengths between them, such that the DNA can fold back upon itself without particular constraints. Larger distances may be adjusted so as to modulate the chances of different enhancers to encounter their cognate promoter – a model inspired by Grosveld's model for the transcriptional regulation of the human β-globin gene complex (Fraser & Grosveld, 1998; Gribnau et al., 1998; Trimborn et al., 1999); inspired also by the analysis of heterochromatinization of euchromatic sequence repeats as a function in part of the distance between the repeats and constitutive heterochromatin. The latter relationship is illustrated with particular relief by a dominant mutation of the *Drosophila brown* gene (bw^D) (Csink & Henikoff, 1996; Dernburg et al., 1996; Csink & Henikoff, 1998a; Marshall & Sedat, 1999).

The looping models had been preceded – but now are also followed – by models of structural propagation of a certain type of regular high-order chromatin structure. These models come in three versions: the procession of macromolecules along the DNA, the directional accretion of certain factors present in the nuclear sap (a kind of linear crystallization), and switches in the intermolecular interaction pattern, in a domino effect, of factors already bound (Zuckerkandl, 1974; Stubblefield, 1986; Zuckerkandl, 1990; Renauld et al., 1993; Zuckerkandl & Hennig, 1995; Hecht, Strahl-Bolsinger & Grunstein, 1996; Morcillo et al., 1997; Bulger & Groudine, 1999; Kelley & Kuroda, 2000). The second and third alternatives for the process might apply jointly.

Whichever mode of interaction pertains, the distances between enhancers and promoters might come into play in regard to the developmental time it takes to obtain the regulatory effects, but not in regard to the structural stability of the resulting local complexes. It is possible that such a time factor is of observable magnitude, in which case greater average distances between enhancers and promoters would result in a slowing of development and vice-versa – a matter to be investigated. If such a relationship obtained, then, potentially, the observed correlation between slowdowns of development and increases in genome size (Cavalier-Smith, 1978) could in part be caused, not by genome size as such, but by the mean distances

between enhancers and promoters. The suggestion is supported by the observation that 'there is a negative correlation between the developmental rate and genome size even when nuclear and cytoplasmic volumes are controlled for, indicating that developmental rate is related to genome size, not just to nuclear volume' (Gregory, 2001).

Tether DNA may play a number of important roles in eukaryotes. Again, the 'junk' would have developmental implications and could thus become selectable, not much, in this respect, as a sequence, but as a distance – or become functional without being selected. Insertions into sequences used as tethers may usually be fixed by genetic drift, but under some circumstances their cumulative effect probably is selected for or against (cf. Knight & Ackerly, in press).

Functions of nucleotide pluralities of the order of 1000 kb

It was mentioned that similar factors, by binding to different sequences, may recruit these sequences for similar functions. At a polynucleotide scale one order of magnitude higher than the one previously considered, namely, 1000 kb, another illustration of this observation is furnished by sequences capable of forming centromeres. Factor specificity, here, is compatible with low functional density of DNA. Long sequences of low functional density and equivalent function can be so different that sequence alignments often will not detect homology above the significance level. Thus, centromeric DNA sequences bear little or no similarity across species (Karpen & Allshire, 1997). Miklos (1985), in his classical article on 'localized highly repetitive DNA,' guided by the 'overwhelming variation at all levels,' could not escape from formulating invalid general inferences regarding functions.

Centromere formation is not necessarily inducible *in vitro* from naked DNA when it is inducible from DNA that is already in the form of chromatin (Willard, 2001). There are sequences, however, that are particularly fit for centromere formation, notably, in humans, α-satellite. In the presence of α-satellite and of the required factors, centromere formation can be induced even from naked DNA. Yet, under cellular conditions, these particularly favorable sequences are neither necessary nor sufficient for centromere formation (Karpen & Allshire, 1997; Murphy & Karpen, 1998; Maggert & Karpen, 2000).

It is important in this regard to consider the formation of neocentromeres. They can be induced in sequences that do not share with most centromeres typical sequence features such as the presence of satellite DNA, nor contain any other obvious diagnostic sequence features (see Willard, 2001). A noncentromeric chromatin region can be turned into a neocentromere without any sequence change having occurred (Barry et al., 2000). Some general sequence conditions for this propensity have been demonstrated, in particular, the presence of flanking heterochromatin both 5′ and 3′ (Henikoff, Ahmad & Malik, 2001) and other traits of long-range organization (Gindullis et al., 2001). Except for such constraints, expressed in some general, probably periodic (Zinkowski, Meyne & Brinkley, 1991) features of DNA, most of the prerequisites for the formation of the DNA-associated complexes are presumably contained in the structures of a set of proteins – among which nucleosomal histone H3-like proteins are prominent (Choo, 2000; Henikoff et al., 2000). It seems that one condition for a sequence to have the potential to form a centromere is the presence of a repeat structure such that the free energy of formation of single-stranded ordered DNA structures is favorable relative to that of the Watson–Crick duplex (Catasti et al., 1999). The high-order structure of the centromere presumably forms cooperatively, and once formed, is stable and heritable (Maggert & Karpen, 2000). The protein factors are better conserved across species than the DNA sequence (Dobie et al., 1999).

DuPraw (see Csink & Henikoff, 1998b) and Csink and Henikoff (1998b) proposed that centromeres are formed at the site of the last region of a chromosome to replicate, by virtue of this region being the last. In this model, as in the model suggested here, centromere formation depends more strongly on protein factors than on DNA sequence.

Before a centromere is established over a particular array of sequences, its functional density, as centromere, is zero. Subsequent to neocentromere formation, the functional density of the DNA involved has the value conferred upon it by the number of contacts per sequence length that the factors establish with distinct nucleotides. Thus, at a certain level of nucleotide plurality, functional density of a segment of DNA need not be to any large extent an intrinsic property of the DNA sequence, but can principally be defined by sets of protein factors capable of interacting with sectors of DNA that had not been suspected to have the corresponding structural wherewithal. Touch

the junk with the right mix of factors and you get function, just as, in the desert, Moses got water to flow out of a rock by hitting it with a stick.

Thus, at the 1000 kb level, in higher organisms, neocentromeres furnish the prototype of a process that amounts to a transposition of functions without transposition of sequences. It takes a large amount of noncoding sequences to offer such opportunities.

Functions of nucleotide pluralities of an order of more than 1000 kb

Yet higher functional units of nucleotide counts, higher than the order of 1000 kb, will be considered very briefly. They seem to exist as functional units in relation to both effective mechanics of chromosome behavior (e.g., Csink & Henikoff, 1998a) and to gene regulation in larger groups of genes (e.g., Patterton & Wolffe, 1996). The regional chromosomal positions of genes may correlate with the specific dynamic behavior of parts of chromosomes relative to the nuclear lamina, to the nuclear matrix, to cell polarity, and to other characters of nuclear fine structure, many of which may have changing relations with parts of chromosomes as a function of development and of the cell cycle phase (Dernburg et al., 1996; Csink & Henikoff, 1998a; Brown, 1999; Marshall & Sedat, 1999).

Of direct relevance here are the existence of different nuclear compartments, namely, the chromosomal territories and interchromatin compartment (Schul, de Jong & Driel, 1998; Cremer & Cremer, 2001), and the fact that transcription and other processes occur in specialized and localized multiple 'factories,' in each of which a particular type of RNA polymerase is bound to the nuclear matrix (Jackson, 1997). Even as they remain aligned in an established order along the chromosome, genes no doubt often need to be able to switch individually or in small groups from one nuclear compartment to another. Therefore, tethers or 'leashes' between genes and groups of genes, again, are probably used to provide a requisite independence of movement. An illustration of such independence is furnished by the pairing of euchromatic repeat sequences with heterochromatin (Henikoff, 1996), a process whose effects have been considered here as having been turned by evolution into regulatory functions. Motions independent from those of whole chromosomes or chromosome arms would frequently seem indispensable when topographical changes in intergenic relations are to be enacted, and when associations of specific chromosomal regions with the nuclear periphery or the nuclear interior are to be switched. For functional processes involving motion a fair amount of 'junk DNA' may be consumed.

Obviously, a sequence, short or long, can function as a tether in one respect and, at the same time, within the tether's boundaries, fill functional roles in other respects. Regarding the human β-globin locus, for example, it seems that a localisation away from centromeric heterochromatin is required to achieve general hyperacetylation of histones H3 and H4 and thus an open chromatin structure of the locus (Schübeler et al., 2000). Here, the whole length of DNA between the closest block of heterochromatin and the globin gene complex plays the role of a tether, though many other genes probably populate this tether – this time a tether in charge of keeping a gene complex at arms' length relative to a repressing zone of the genome. Effects of this kind require genomes of sizes far beyond the sum of their coding sequences. The very existence if heterochromatin – which is functional (e.g., Zuckerkandl & Hennig, 1995) – contributes far beyond its own boundaries to generating a c-value paradox in the haploid genome.

Although chromosomal regions of increased and characteristic GC content (isochores) might often not exceed the magnitude of 1000 kb nucleotide plurality (Bernardi, 1989), there are larger-scale subdivisions in genomes whose built-in features (Bernardi, 1989, 1995; Holmquist, 1989, 1992) almost guarantee functional significance, at the very least through affecting the evolution of function. Holmquist (1992, 1994) presented a compelling picture of self-organization and self-directed evolution of genome compartments on a 10,000 kb scale, namely, on the scale of metaphase bands.

It may be that corresponding large-scale subdivisions are expressed in sperm through the distribution of sequences (Schmid, 1998) over the nucleoprotamine and the nucleohistone compartments. T-bands in human metaphase chromosomes, which are particularly GC-rich and SINE-rich, represent 15% of all bands and contain 65% of the genes (Holmquist, 1992). So does the SINE-rich nucleohistone fraction of sperm DNA represent 15% of the genome (Gatewood et al., 1987, 1990). The coincidence of these figures raises the question whether they might generally relate to the same fractions of DNA.

Not addressed here is an epigenetic effect extending over vast regions of chromatin, namely, dosage compensation. Hennig (1999) and Kelley and

Kuroda (2000) propose chromatin-associated nucleoprotein complexes as the source of dosage compensation in X-chromosomes.

At the top of the hierarchy of sizes of functional DNA units, namely, in regard to a function of the genome mass as a whole, the functional density of all sequences is zero. Sequence is no longer involved in the DNA's effect, only amount is (Cavalier-Smith & Beaton, 1999). In the widest sense, whether 'personally' functional, parafunctional, or persistently nonfunctional, every DNA sequence could of course be regarded as minutely functional if it participated in functional effects of genome size.

Genome size has been found to correlate with nuclear size, cell size, cell numbers, rate of cell division, rate of development, metabolic rate, and developmental end points, including neural complexity (Cavalier-Smith, 1982; Roth, Blanke & Wake, 1994; Roth, Nishikawa & Wake, 1997; Gregory & Hebert, 1999; Gregory, 2001; see Box 1, Petrov, 2001). Extreme environmental conditions correlate with smaller genome sizes (Knight & Ackerly, 2002). Weeds offer a telling example of the correlation between small c-value and small nuclear DNA content on the one hand and rapid development on the other (Bennett, Leitch & Hanson, 1998). Whichever way causal relationships go – they might well be circular to a significant extent – the correlations have been shown to have wide functional implications. That implications and consequences of this kind would routinely escape natural selection would seem unnatural. The possible regulatory effect of varying inter- and intragenic distances needs, however, to be assessed before the additional contributions of total genomic mass as such can be definitively evaluated.

Selection and election

As has already been emphasized, there seems to be a general lesson to be learned from the centromere and, more generally, from heterochromatin, one probably to be applied to large parts of the genomes of higher organisms: for high-nucleotide-plurality functions, selection plays primarily on the factors (proteins and surely also a large number of RNAs). Beyond selection comes election: the factors can elect a preexisting DNA sequence to use as their nest.

Clearly, the factors cannot have evolved collectively in preparation for being able to take advantage of new nesting opportunities. But their proclivity to build up their complexes primarily through mutual interaction left a great many degrees of freedom to the sequence of their DNA partner. Hence, the 'overwhelming' variation in noncoding sequences compatible with genomic functions. Hence, also, the irritation if not scorn manifested by many at the mere suggestion that such sequences may 'make sense.' In reality, when it comes to reading sequences, we are still mostly illiterate. The factors aren't. And their avocation is mostly to read each other.

At certain levels of nucleotide plurality, many noncoding sequences are then probably not outright meaningless, but are only conditionally meaningful (functional). These conditionally meaningful – and thus also conditionally meaningless – long sequences – meaningful or meaningless at the particular level of nucleotide plurality under consideration – can contain a number of components that are meaningful at lower scales of sequence lengths; alternatively, they can themselves represent components of yet longer sequences that are meaningful at an even more inclusive scale. The genome is thus a mosaic of the functional and the nonfunctional at different levels of nucleotide plurality, with the nonfunctional neighboring the functional at any given level, and with sums of the functional and nonfunctional sequences being functional again in a different way at a higher level of nucleotide plurality.

As widely known, trash can be elected to office. In genomes, the process has, implicitly, been deemed applicable only to components of punctate gene regulation. Election to sectorial functions is likely, however, to represent a further important feature of eukaryote evolution.

If one minute of philosophy were (again) permitted, this minute should be used here to remember certain statements, made in February 2001, when the success of the Human Genome Project was being celebrated. At that occasion, several scientists proclaimed that the genome sequence represented 'the script of life.' If the aim is indeed to be able to 'read' living systems, knowing the correct alignment of nucleotides is certainly a condition. Yet, by itself this alignment of nucleotides cannot provide even a partial reading, only the shadow of a partial reading. It cannot document the unfolding of interactions among proteins and among proteins and polynucleotides nor account for the coordination in space and time of events whereby such interactions are brought about, distributed, shut down, and modified. Investigating these relationships would bring us closer to discovering a script of life if they represented a script. But we don't call script a multi-

dimensional interaction pattern that unwinds in time and space. The nucleotide sequence is but an organized collection of fragments of this system. Within the nucleotide sequence itself, the genetic code is only one of the codes, as Trifonov (1989, 1996) realized. Beside the triplet code, there is, for example, a 'transcription code,' a gene splicing code, a translation pausing code, a 'DNA structure code,' a 'chromatin code,' a 'translation framing code,' etc. (Trifonov, 1996). Yet, the 'meaning' of DNA sequences is only fractionally revealed by DNA sequence codes. All codes are interacting within a 'code of conduct' at the scale of the genome, a sort of code of higher order. DNA sequences by themselves alone contain no trace of a code of conduct. Much of the information that translates into specifically patterned effects is scattered in biological space. Assuredly, the amount of 'intelligence' that one can gather from just a fraction of a code is limited. That is how most of the 'script of life' can afford to look so dumb.

c-value paradox and complexity

I would be reluctant to join Cavalier-Smith in calling all DNA that does not function under a high-functional-density regimen 'secondary.' The molecular form of present political correctness may still yield to a Declaration of Independence worked out by a Committee of Noncoding Sequences and that holds all sequences to be created equal. One area where prejudice against the bulk of non-protein-coding sequences is rampant relates to their contribution to complexity.

It has indeed often been proclaimed (e.g., Cavalier-Smith & Beaton, 1999) that DNA sequences other than protein-coding sequences and their immediate regulatory noncoding dependencies have no relevance to an organism's complexity. This is incorrect in regard to the haploid genome, because (1) DNA components of regulatory systems have multiplied and spread over the genomes of higher organisms, with their mutual distances and topological relationships often endowed with phenotypic effects. The link between regulatory complexity and the availability of increased amounts of non-protein-coding DNA is perceptible directly through the frequent presence of multiple enhancers per gene. Ensuing functional requirements in terms of DNA tethers and other sequence implements contribute to the 'need' of 'junk DNA,' if higher regulatory complexity is to be attained and maintained; (2) SINEs and LINEs formerly treated as purely selfish can now be considered on experimental grounds to be likely to fill important functions as genes. If such is the case, there is no reason to postulate that SINEs and LINEs, unlike other genes, fail to contribute to gene interaction complexity and organismal complexity; (3) the epigenetic functional processes present in the nucleus of the eukaryote cell at multiple levels of nucleotide integration qualify as contributors to the complexity of the gene interaction system and thus to the complexity of the organism; (4) these additional types of functions – additional to those known from the prokaryotes – are in fact likely to have paved the way for more complex organisms to evolve by introducing additional mechanisms of genic, chromosomal, and cellular control; furthermore, (5) epigenetic gene control is based not only on middle-repetitive sequences, it is also linked to interacting tandem repeats of oligonucleotides – namely, minisatellites and microsatellites (e.g., Gilmour et al., 1989; Catasti et al., 1999). Hence an increase in opportunities for differential gene and chromosomal regulation (e.g., the involvement of TTAGGG repeats in the control of telomeres). Tandem repeats, whether single-stranded, double helix or triple helix structures, and whether involved in gene or chromosome control, further extend the variety of available recognition motifs for proteins and protein complexes, and thus clearly contribute to the complexity of the regulatory system. (See note added in proof.)

Finally, (6), when increases in c-value become extreme as they do in salamanders (Roth, Blanke & Wake, 1994, Roth, Nishikawa & Wake, 1997), complexity – manifestly, in this case, organismal complexity – is again affected, though in the opposite direction: because of developmental restrictions, it decreases significantly.

The present focus was on the range of nucleotide plurality where the direct (rather than the inverse) correlation between genome size and complexity applies, namely, in particular, on the range connected to what has been characterized here as c-value paradox I. Even though complexity increases within this range, it cannot be expected to increase in linear proportion to c-value, because the number of additional functions decreases as the number of nucleotides per function increases. Admittedly, it has not been determined how complexity increases linked to higher nucleotide pluralities may be measured. Nevertheless, a high nucleotide consumption by additional functions is the reason why, even before the role of total nuclear DNA content is brought into

the picture, the paradox of c-values is found to be reduced.

Epigenetic control as part of a basic uniformity of living systems

That living systems are unitary in all their incarnations has long since been clear, especially since the discovery of the universal genetic code; but unitary in what ways throughout the various hierarchical levels of biological integration? This takes more effort to explore. It was soon realized (e.g., Zuckerkandl, 1975) that genes – at least their component parts representing protein domains – are in general extremely ancient objects through which all living things, including different kingdoms of organisms, are closely related. Not present as a concept until the early eightees was the extraordinary antiquity and continued conservation of a number of basic regulatory relationships among genes, relationships about as old as the genes themselves (Zuckerkandl, 1983, 1994). Extraordinary antiquity is also the hallmark, under one guise or another, of the epigenetic control of DNA function, even though the 'sectorial' versions of this control might be limited to the eukaryotes. All bioevolution that is not cut short seems likely, eventually, to feature genomes whose sequences can assume alternate structures with alternate regulatory effects.

There is no single solution to the c-value paradox, but a partial one is indeed to be found in the mechanisms of epigenetic control which in eukaryotes depend upon a considerable supply of non-protein-coding sequences.

At the same time, no functional aspect of the system could presumably arise and be maintained without the essential contribution of the genes and their associated high-functional-density non-protein-coding sequences required for punctate regulation. The 'news,' for anyone who may not have believed it yet, is that the presence of large amounts of lower- and low-functional-density noncoding sequences is a prerequisite for the existence of the most complex organisms.

Note added in proof

Conserved non-protein-coding, yet transcribed DNA and their corresponding RNAs represent the 'genomic dark matter' (Greg Hannon, in a seminar at Stanford, November 29, 2001), upon which light is now shed as the RNAs are being integrated into regulatory systems, transcriptional and translational alike (Bernstein, Denli & Hannon, 2001; Mattick & Gagen, 2001). A large population of diverse noncoding RNAs promises to contribute in essential ways to the complexity of developing and mature gene interaction systems and organisms.

Acknowledgements

My thanks, first and foremost, to Dmitri Petrov who elicited this paper, encouraged the protracted process of its generation, and, through his critical input, much improved the analyses presented. A great many thanks for their input also to Jean-Maurice Dura, Steve Henikoff, Aaron Hirsh, Charley Knight, and Carl Schmid. I am especially grateful to Allan Campbell in the Department of Biological Sciences at Stanford for the hospitality of his laboratory.

References

Ahmad, K. & S. Henikoff, 2001. Modulation of a transcription factor counteracts heterochromatic gene silencing in *Drosophila*. Cell 104: 839–847.

Ainscough, J.F., R.M. John & M.A. Surani, 1998. Mechanism of imprinting on mouse distal chromosome 7. Genet. Res. 72: 237–245.

Ainscough, J.F., L. Dandolo & M.A. Surani, 2000. Appropriate expression of the mouse H19 gene utilises three or more distinct enhancer regions spread over more than 130 kb. Mech. Dev. 91: 365–368.

Bailey, J.A., L. Carrel, A. Chakravarti & E.E. Eichler, 2000. Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. Proc. Natl. Acad. Sci. USA 97: 6634–6639.

Barry, A.E., M. Bateman, E.V. Howman, M.R. Cancilla, K.M. Tainton, D.V. Irvine, R. Saffery & K.H. Choo, 2000. The 10q25 neocentromere and its inactive progenitor have identical primary nucleotide sequence: further evidence for epigenetic modification. Genome Res. 10: 832–838.

Beckman, J.S., V. Brendel & E.N. Trifonov, 1986. Intervening sequences exhibit distinct vocabulary. J. Biomol. Struct. Dyn. 4: 391–400.

Beckman, J.S. & E.N. Trifonov, 1991. Splice junctions follow a 205-vase ladder. Proc. Natl. Acad. Sci. USA 88: 2380–2383.

Bell, A.C., A.G. West & G. Felsenfeld, 2001. Insulators and boundaries: versatile regulatory elements in the eukaryotic genome. Science 291: 447–450.

Bennett, M.D., I.J. Leitch & L. Hanson, 1998. DNA amounts in two samples of angiosperm weeds. Ann. Bot. 82: 121–134.

Bensasson, D., D.A. Petrov, D.X. Zhang, D.L. Hartl & G.M. Hewitt, 2001. Genomic gigantism: DNA loss is slow in mountain grasshoppers. Mol. Biol. Evol. 18: 246–253.

Berezovsky, I.N. & E.N. Trifonov, 2001. Van der Waals locks: loop-n-lock structure of globular proteins. J. Mol. Biol. 307: 1419–1426.

Bergman C.M. & M. Kreitman, 2001. Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. Genome Res. 11: 1319–1320.

Bernardi, G., 1989. The isochore organization of the human genome. Annu. Rev. Genet. 23: 637–661.

Bernardi, G., 1995. The human genome: organization and evolutionary history. Annu. Rev. Genet. 29: 445–476.

Bernstein, E., A.M. Denli & G.J. Hannon, 2001. The rest is silence. RNA 7: 1509–1521.

Boeke, J.D., 1997. LINES and Alus – the polyA connection. Nature Genet. !6: 6–7.

Bolshoy, A., K. Shapiro, E.N. Trifonov & I. Ioshikhes, 1997. Enhancement of the nucleosomal pattern in sequences of lower complexity. Nucl. Acids Res. 25: 3248–3254.

Brini, A.T., G.M. Lee & J.-P. Kinet, 1993, Involvement of Alu sequences in the cell-specific regulation of transcription of the γ chain of Fc and T cell receptors. J. Biol. Chem. 268: 1355–1361.

Britten, R.J., 1994. Evolutionary selection against change in many Alu repeat sequences interspersed through primate genomes. Proc. Natl. Acad. Sci. USA 91: 5992–5996.

Britten, R.J., 1995. Quantitative study of Alu repeated sequences in primate genomes, yielding insight into their sources and evolution, pp. 223–232 in The Impact of Short Interspersed Elements (SINEs) on the Host Genome, edited by R.J. Maraia. Landes Company.

Britten, R.J., 1996a. DNA sequence insertion and evolutionary variation in gene regulation. Proc. Natl. Acad. Sci. USA 93: 9374–9377.

Britten, R.J., 1996b. Cases of ancient mobile element DNA insertions that now affect gene regulation. Mol. Phylogenet. Evol. 5: 13–17.

Britten, R.J., 1997. Mobile elements inserted in the distant past have taken on important functions. Gene 205: 177–182.

Brown, K., 1999. Nuclear structure, gene expression and development. Crit. Rev. Eukaryot Gene Expr. 9: 203–212.

Bulger, M. & M. Groudine, 1999. Looping versus linking. Genes Dev. 13: 2467–2477.

Burns, J.L., D.A. Jackson & A.B. Hassan, 2001. A view through the clouds of imprinting. FASEB J. 15: 1694–1703.

Caplan, A.I. & C.P. Ordahl, 1978. Irreversible gene expression model for control of development. Science 201: 120–130.

Casares, F., W. Bender, J. Merriam & E. Sanchez-Herrero, 1997. Interactions of *Drosophila* Ultrabithorax regulatory regions with native and foreign promoters. Genetics 145: 123–137.

Casavant, N.C., L. Scott, M.A. Cantrell, L.E. Wiggins, R.J. Baker & H.A. Wichman, 2000. The end of the LINE? Lack of recent L1 activity in a group of South American rodents. Genetics 154: 1809–1817.

Catasti, P., X. Chen, S.V. Mariappan, E.M. Bradbury & G. Gupta, 1999. DNA repeats in the human genome. Genetica 106: 15–36.

Cavalier-Smith, T., 1978. 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.

Cavalier-Smith, T., 1982. Skeletal DNA and the evolution of genome size. Annu. Rev. Biophys. Bioeng. 11: 273–302.

Cavalier-Smith, T. & M.J. Beaton, 1999. The skeletal function of non-genic nuclear DNA: new evidence from ancient cell chimaeras. Genetica 106: 3–13.

Chan, C.-S., L. Rastelli & V. Pirottta, 1994. A polycomb response element in the Ubx gene that determines an epigenetically inherited state of repression. EMBO J. 13: 2553–2564.

Cho, G. & R.F. Doolittle, 1997. Intron distribution in ancient paralogs supports random insertion and not random loss. J. Mol. Evol. 44: 573–584.

Choo, K.H.A., 2000. Centromerization. Trends Cell Biol. 10: 182–188.

Chu, W.M., R. Ballard, B.W. Carpick, B.R. Williams, C.W. Schmid, 1998. Potential Alu function:regulation of the activity of double-stranded RNA-activated kinase PKR. Mol. Cell Biol. 18: 58–68.

Clemens, M.J., 1996. Protein kinases that phosphorylate eIF2 an eIF2B, and their role in eukaryotic cell translational control, pp. 139–172. in Translational Control, edited by J.W.B. Hershey, M.B. Mathews & N. Sonenberg, Cold Spring Harbor Laboratory Press, Plainview, NY.

Cremer, T. & C. Cremer, 2001. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nature Rev. 2: 292–301.

Csink, A.K. & S. Henikoff, 1996. Genetic modification of heterochromatic association and nuclear organization in *Drosophila.* Nature 381: 529–531.

Csink, A.K. & S. Henikoff, 1998a. Large-scale chromosomal movements during interphase progression in *Drosophila*. J. Cell Biol. 143: 13–22.

Csink, A.K. & S. Henikoff, 1998b. Something from nothing: the evolution and utility of satellite repeats. Trends Genet. 14: 200–204.

Denisov, D.A., E.S. Shpigelman & E.N. Trifonov, 1997. Protective nucleosome centering at splice sites as suggested by sequence-directed mapping of the nucleosomes. Gene 205 (1–2): 145–149.

Dernburg, A.F., K.W. Broman, J.C. Fung, W.F. Marshall, J. Philips, D.A. Agard & J.W. Sedat, 1996. Perturbation of nuclear architecture by long-distance chromosome interactions. Cell 85: 745–759.

de Souza. S.J., M. Long, R.J. Klein, S. Roy, S. Lin & W. Gilbert, 1998. Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. Proc. Natl. Acad. Sci. USA 95: 5094–5099.

Deutsch, M. & M. Long, 1999. Intron-exon structures of eukaryotic model organisms. Nucl. Acids Res. 27: 3219–3228.

DiLeone, R.J., G.A. Marcus, M.D. Johnson & D.M. Kingsley, 2000. Efficient studies of long-distance Bmp5 gene regulation using bacterial artificial chromosomes. Proc. Natl. Acad. Sci. USA 97: 1612–1617.

Dobie, K.W., K.L. Hari, K.A. Maggert & G.H. Karpen, 1999. Centromere proteins and chromosome inheritance: a complex affaire. Curr. Opin. Genet. Dev. 9: 206–217.

Doolittle, R.F., D.F. Feng, M.S. Johnson & M.A. McClure, 1989. Origins and evolutionary relationships of retroviruses. Quart. Rev. Biol. 64: 1–30.

Doolittle, W.F. & C. Sapienza, 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature 284: 601–603.

Dorer D.R. & S. Henikoff, 1994. Expansion of transgene repeats cause heterochromatin formation and gene silancing in *Drosophila*. Cell 77: 993–1002.

Dorer, D.R. & S. Henikoff, 1997. Transgene repeat arrays interact with distance heterochromatin and cause silencing in *cis and trans.* Genetics 147: 1181–1190.

Duboule, D., 1992. The vertebrate limb: a model system to study the *Hox*/HOM gene network during development and evolution. BioEssays 14: 375–384.

Duret, L., D. Mouchiroud & C. Gautier, 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J. Mol. Evol. 40: 308–317.

Eickbush, T.H., 1992. Transposing without ends: the non-LTR retrotransposable elements. New Biol. 4: 430–440 .

Eissenberg, J.C. & A.J. Hilliker, 2000. Versatility of conviction: heterochromatin as both a repressor and an activator of transcription. Genetica 109: 19–24.

Englander, E.W., A.P. Wolffe & B.H. Howard, 1993. Nucleosome interactions with a human Alu element. J. Biol. Chem. 268: 19565–19573.

Fanti, L., D.R. Dorer, M. Berloco, S. Henikoff & S. Pimpinelli, 1998. Heterochromatin protein 1 binds transgene arrays. Chromosoma 107: 286–292.

Felsenfeld, G., 1992. Chromatin as an essential part of the transcriptional mechanism. Nature 355: 219–224.

Felsenfeld, G., J. Boyce, J. Chung, D. Clark & V. Studitsky, 1996. Chromatin structure and gene expression. Proc. Natl. Acad. Sci. USA 93: 9384–9388.

Ferguson-Smith, A.C. & M.A. Surani, 2001. Imprinting and the epigenetic asymmetry between parental genomes. Science 293: 1086–1089.

Fiering, S., E. Whitelaw & D.I.K. Martin, 2000. To be or not to be active: the stochastic nature of enhancer action. BioEssays 22: 381–387.

Francastel, C., M.C. Walters, M. Groudine & D.L. Martin, 1999. A functional enhancer suppresses silencing of a transgene and prevents its localization close to centromeric heterochromatin. Cell 99: 259–269.

Fraser, P. & F. Grosveld, 1998. Locus control regions, chromatin activation and transcription. Curr. Opin. Cell Biol. 10: 361–365.

Furano, A.V., 2000. The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. Prog. Nucl. Acid Res. Mol. Biol. 64: 255–294.

Gatewood, J.M., G.R. Cook, R. Balhorn, E.M. Bradbury & C.W. Scmid, 1987. Sequence-specific packaging of DNA in human sperm chromatin. Science 236: 962–964.

Gatewood, G.M., G.R. Cook, R. Balhorn, C.W. Scmid & A.M. Bradbury, 1990. Isolation of four core histones from human sperm chromatin representing a minor subset of somatic histones. J. Biol. Chem. 265: 20662–20666.

Gaunt, S.J. & P.B. Singh, 1990. Homeogene expression patterns and chromosomal imprinting. Trends Genetics 6: 208–212.

Gilmour, D.S., G.H. Thomas & S.C. Elgin, 1989. Drosophila nuclear proteins bind to regions of alternating C and T residues in gene promoters. Science 245: 1487–1490.

Gindullis, F., C. Desel, I. Galass & T. Schmidt, 2001. The large-scale organization of the centromeric region in Beta species. Genome Res. 11: 253–265.

Gregory, T.R., 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol. Rev. Camb. Philos. Soc. 76: 65–101.

Gregory, T.R. & P.D. Hebert, 1999. The modulation of DNA content: proximate causes and ultimate consequences. Genome Res. 9: 317–324.

Gribnau, J., E. de Boer, T. Trimborn, M. Wijgerde, E. Milot, F. Grosveld & P. Fraser, 1998. Chromatin interaction mechanism of transcriptional control *in vivo*. EMBO J. 20: 6020–6027.

Hartl, D.L., D.E. Dykhuizen & A.M. Dean, 1985. Limits of adaptation: the evolution of selective neutrality. Genetics 111: 655–674.

Hasse, A. & W.A. Schulz, 1994. Enhancement of reporter gene *de novo* methylation by DNA fragments from the α-fetoprotein control region. J. Biol. Chem. 269: 1821–1826.

Hecht, A., S. Strahl-Bolsinger & M. Grunstein, 1996. Spreading of transcriptional repressor SIR3 from telomeric heterochromatin. Nature 383: 92–96.

Henderson, I.R., P. Owen & J.P. Nataro, 1999. Molecular switches – the ON and OFF of bacterial phase variation. Mol. Microbiol. 33: 919–932.

Hendrich, B.D. & H.F. Willard, 1995. Epigenetic regulation of gene expression: the effect of altered chromatin structure from yeast to mammals. Hum. Mol. Genet. 4 (Spec. No.): 1765–1777.

Henikoff, S., 1996. Position-effect variegation in *Drosophila*: recent progress, pp. 319–334 in Epigenetic Mechanisms of Gene Regulation. Cold Spring Harbor Laboratory Press.

Henikoff, S. & M.A. Matzke, 1997. Exploring and explaining epigenetic effects. Trends Genet. 13: 293–295.

Henikoff S., K. Ahmad, J.S. Platero & B. van Steensel, 2000. Heterochromatic deposition of centromeric histone H3-like proteins. Proc. Natl. Acad. Sci. USA 97: 716–721.

Henikoff, S., K. Ahmad & H.S. Malik, 2001. The centromere paradox: stable inheritance with rapidly evolving DNA. Science 293 (5532) 1098–1102.

Hennig, W., 1999. Heterochromatin. Chromosoma 108: 1–9.

Herzel, H., O. Weiss & E.N. Trifonov, 1999. 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding. Bioinformatics 15: 187–193.

Hirsh, A.E. & H.B. Fraser, 2001. Protein dispensability and rate of evolution. Nature 411: 1046–1049.

Holmquist, G.P., 1989. Evolution of chromosome bands: molecular ecology of noncoding DNA. J. Mol. Evol. 28: 469–486.

Holmquist, G.P., 1992. Chromosome bands, their chromatin flavors, and their functional features. Am. J. Hum. Genet. 51: 17–37.

Holmquist, G.P., 1994. Chromatin self-organization by mutation bias. J. Mol. Evol. 39: 436–438.

Horard B., C. Tatout, S. Poux & V. Pirrotta, 2000. Structure of a polycomb response element and *in vitro* binding of polycomb group complexes containing GAGA factor. Mol. Cell. Biol. 20: 3187–3197.

Humphrey, G.W., E.W. Englander & B.H. Howard, 1996. Specific binding sites for Pol III transcriptional repressor and Pol II transcription factor YY1 within the internucleosomal spacer region in primate Alu repetitive elements. Gene Express. 6: 151–168.

Ioshikhes, I., E.N. Trifonov & M.Q. Zhang, 1999. Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. Proc. Natl. Acad. Sci. USA 96 (6): 2891–2895.

Ishihara, K., N. Hatano, H. Furuumi, R. Kato, T. Iwaki, K. Miura, Y. Jinno & H. Sasaki, 2000. Comparative genomic sequencing identifies novel tissue-specific enhancers and sequence elements for methylation-sensitive factors implicated in Igf2/H19 imprinting. Genome Res. 10: 664–671.

Jackson, D.A., 1997. Chromatin domains and nuclear compartments: establishing sites of gene expression in eukaryotic nuclei. Mol. Biol. Rep. 24: 209–220.

Jurka, J., 1995. Origin and evolution of Alu repetitive elements, pp. 25–41 in The Impact of Short Interspersed Elements (SINEs) on the Host Genome, edited by R.J. Maraia. Springer Verlag, Heidelberg.

Jurka, J. & E. Zuckerkandl, 1991. Free left arms as precursor molecules in the evolution of Alu sequences. J. Mol. Evol. 33: 49–56.

Kariya, Y., K. Kato, Y. Hayashizaki, S. Himeno, S. Tarui & K. Matsubara, 1987. Revision of consensus sequence of human Alu repeats – a review. Gene 53: 1–10.

Karpen, G.H., M.H. Le & H. Le, 1996. Centric heterochromatin and the efficiency of achiasmate disjunction in *Drosophila* female meiosis. Science 273: 118–122.

Karpen, G.H. & R.C. Allshire, 1997. The case for epigenetic effects on centromere identity and function. Trends Genet. 13: 489–496.

Kassis, J.A., M.L. Wong & P.H. O'Farrell, 1985. Electron microscopic heteroduplex mapping identifies regions of the engrailed locus that are conserved between *Drosophila melanogaster* and *Drosophila virilis*. Mol. Cell. Biol. 5: 3600–3609.

Kassis, J.A., C. Desplan, D.K. Wright & P.H. O'Farrell, 1989. Evolutionary conservation of homeodomain-binding sites and other sequences upstream and within the major transcription unit of the *Drosophila* segmentation gene engrailed. Mol. Cell. Biol. 9: 4304–4311.

Kelley, R.L. & M.I. Kuroda, 2000. Noncoding RNA genes in dosage compensation and imprinting. Cell 103: 9–12.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

Kimura, R.H., P.V. Choudary & C.W. Schmid, 1999. Silk worm Bm1 RA increases following cellular insults. Nucl. Acids Res. 27: 3380–3387.

Kimura, R.H., P.V. Choudary, K.K. Stone & C.W. Schmid, 2001. Stress induction of Bm1 RNA in silk worm larvae: SINES, an unusual class of stress genes. Cell Stress Chaper. 6: 263–272.

Knight, C.A. & D.D. Ackerly, 2002. Variation in nuclear DNA content across environmental gradients: a quantile regression analysis. Ecology Lett. (in press).

Koop, B.F., 1995. Human and rodent DNA comparisons: a mosaic model of genomic evolution. Trends Genet. 11: 367–371.

Koop, B.F., Hood, L., 1994. Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. Nat. Genet. 7: 48–53.

Kuo, K.W., H.M. Shen, Y.S. Huang & W.C. Leung, 1998. Expression of transposon LINE-1 is relatively human-specific and function of the transcript may be proliferation-essential. Biochem. Biophys. Res. Commun. 253: 566–570.

Kuska, B., 1998. Should scientists scrap the notion of junk DNA? J. Natl. Cancer Inst. 90: 1032–1033.

Levitzky, V.G., O.A. Podkolodnaya, N.A. Kolchanov & N.L. Podkolodny, 2001a. Nucleosome formation potential of exons, introns, and Alu repeats. Bioinformatics 17: 1062–1064.

Levitsky, V.G., O.A. Podkolodnaya, N.A. Kolchanov & N.L. Podkolodny, 2001b. Nucleosome formation potential of eukaryotic DNA: calculation and promoter analysis. Bioinformatics 17: 998–1010.

Lewis, E.B., 1950. The phenomenon of position effect. Adv. Genet. 3: 73–115.

Lewis, E.B., 1978. A gene complex controlling segmentation in *Drosophila*. Nature 276: 565–570.

Li, W.H. & D. Graur, 1991. Fundamentals of Molecular Evolution. Sinauer, Sunderland, MA.

Li, T., J. Spearow, C.M. Rubin & C.W. Schmid, 1999. Physiological stresses increase mouse short interspersed element (SINE) RNA expression *in vivo*. Gene 239: 367–372.

Li, X., D. Liu & C. Liang, 2001. Beyond the locus control region: new light on beta-globin locus regulation. Int. J. Biochem. Cell Biol. 33: 914–923.

Liu, K., E.P. Sandgren, R.D. Palmiter & A. Stein, 1995. Rat growth hormone gene introns stimulate nucleosome alignment *in vitro* and in transgenic mice. Proc. Natl. Acad. Sci. USA 92: 7724–7728.

Liu, W.M., R.J. Maraia, C.M. Rubin & C.W. Schmid, 1994. Alu transcripts: cytoplasmic localisation and reglation by DNA methylation. Nucl. Acids Res. 22: 1087–1095.

Lloyd, V.K., D.A. Sinclair & T.A. Grigliatti, 1999. Genomic imprinting and position-effect variegation in *Drosophila melanogaster*. Genetics 151: 1503–1516.

Lyon, M.F., 2000. LINE-1 elements and X chromosome inactivation: a function of junk DNA? Proc. Natl. Acad. Sci. USA 97: 6248–6249.

Ma, X., D. Yuan, K. Diepold, T. Scarborough & J. Ma, 1996. The *Drosophila* morphogenetic protein Bicoid binds DNA cooperatively. Development 122: 1195–1206.

Maggert, K.A. & G.H. Karpen, 2000. Acquisition and metastability of centromere identity and function: sequence analysis of a human neocentromere. Genome Res. 10: 725–728.

Makalowski, W., G.A. Mitchell & D. Labuda, 1994. Alu sequences in the coding regions of mRNA: a source of protein variability. Trends Genet. 10: 188–193.

Malik, H.S., W.D. Burke & T.H. Eickbush, 1999. The age and evolution of non-LTR retrotransposable elements. Mol. Biol. Evol. 16: 793–805.

Malik, H.S. & T.H. Eickbush, 2001. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Res. 11: 1187–1197.

Maraia, R.J. & J. Sarrowa, 1995. Alu-family SINE RNA: Interacting proteins and pathways of expression, pp. 163–196 in The Impact of Short Interspersed Elements (SINEs) on the Host Genome, edited by R.J. Maraia. R.G. Landes Company.

Marshall, W.F. & J.W. Sedat, 1999. Nuclear architecture. Results Probl. Cell Differ. 25: 283–301.

Mattick, J.S. & M.J. Gagen, 2001. The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. Mol. Biol. Evol. 18: 1611–1630.

McGinnis, W., R.L. Garber, J. Wirz, A., Kuroiwa & W.J. Gehring, 1984. A homologous protein-coding sequence in *Drosophila* homeotic genes and its conservation in other metazoans. Cell 37: 403–408.

McGowan, R., R. Campbell, A. Peterson & C. Sapienza, 1989. Cellular mosaicism in the methylation and expression of hemizygous loci in the mouse. Genes Dev. 3: 1669–1676.

Mihaly, J., R.K. Mishra & F. Karch, 1998. A conserved sequence motif in *polycomb*-response elements. Mol. Cell 1: 1065–1066.

Miklos, G.L.G., 1985. Localized highly repetitive DNA sequences in vertebrate and invertebrate genomes, pp. 241–321 in Molecular Evolutionary Genetics, edited by R.J. MacIntyre. Plenum Press, New York.

Miller, D., 2000. Analysis and significance of messenger RNA in human ejaculated spermatozoa. Mol. Reprod. Dev. 56: 259–264.

Morcillo P., C. Rosen, M.K. Baylies & D. Dorsett, 1997. Chip, a widely expressed chromosomal protein required for segmentation and activity of a remote wing margin enhancer in *Drosophila*. Genes Dev. 11: 2729–2740.

Moriyama, E.N., D.A. Petrov & D.L. Hartl, 1998. Genome size and intron size in *Drosophila*. Mol. Biol. Evol. 15: 770–773.

Müller, J., 1995. Transcriptional silencing by the polycomb protein in *Drosophila* embryos. EMBO J. 14: 1209–1220.

Murphy, T.D. & G.H. Karpen, 1998. Centromeres take flight: alpha satellite and the quest for the human centromere. Cell 93: 317–320.

Naora, H. & N.J. Deacon, 1982a. Relationship between the total size of exons and introns in protein-coding genes of higher eukaryotes. Proc. Natl. Acad. Sci. USA 79: 6169–6200.

Naora, H. & N.J. Deacon, 1982b. Clustered genes require extragenic territorial DNA sequences. Differentiation 21: 1–6.

Ohno, S., 1997. Beginning of a notion of junk DNA. Presentation made at meeting on 'Junk DNA: the role and the evolution

of non-coding sequences', G. Bernardi, T. Ohta, G. Macaya, organizers, Guanacaste, Costa Rica.

Okada, N., 1990. Transfer RNA-like structure of the human Alu family: implications of its generation mechanism and possible functions. J. Mol. Evol. 31: 500–510.

Orgel, L.E. & F.H. Crick, 1980. Selfish DNA: the ultimate parasite. Nature 284: 604–607.

Orlando, V. & R. Paro, 1993. Mapping polycomb-repressed domains in the bithorax complex using *in vivo* formaldehyde cross-linked chromatin. Cell 75: 1187–1198.

Paro, R. & J.P. Harte, 1996. The role of polycomb group and trithorax group chromatin complexes in the mainteance of determined cell states, pp. 507–528 in Epigenetic Mechanisms of Gene Regulaytion. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Pardue, M.L., O.N. Danilevskaya, K.L. Traverse & K. Lowenhaupt, 1997. Evolutionary links between telomeres and transposable elements. Genetica 100: 73–84.

Patterton, D. & A.P. Wolffe, 1996. Developmental roles for chromatin and chromosomal structure. Dev. Biol. 173: 2–13.

Petrov, D.A., 2001. Evolution of genome size: new approaches to an old problem. Trends Genet. 17: 23–28.

Petrov, D.A. & D.L. Hartl, 1998. High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. Mol. Biol. Evol. 15: 293–302.

Petrov, D.A., T.A. Sangster, J.S. Johnston, D.L. Hartl & K.L. Shaw, 2000. Evidence for DNA loss as a determinant of genome size. Science 287: 1060–1062.

Pirrotta, V., 1997. Mechanisms in *Drosophila* maintain patterns of gene expression. Trends Genet. 13: 314–318.

Poux, S., D. McCabe & V. Pirrotta, 2001. Recruitment of components of polycomb group chromatin complexes in *Drosophila*. Devel. 128: 75–85.

Quentin, Y., 1994. A master sequence related to a free left Alu monomer (FLAM) at the origin of the B1 family in rodent genomes. Nucl. Acids Res. 22: 2222–2227.

Renauld, H., O.M. Aparicio, P.D. Zierath, B.L. Billington, S.K. Chhablani & E.E. Gottschling, 1993. Silent domains are assembled continuously from the telomere and are defined by promoter distance and strength, and by SIR3 dosage. Genes Dev. 7: 1133–1145.

Rzhetsky, A., F.J. Ayala, L.C. Hsu, C. Chang & A. Yoshida, 1997. Exon/intron structure of aldehyde dehydrogenase genes supports the '*introns-late*' theory. Proc. Natl. Acad. Sci. USA 94: 6820–6825.

Robertson, G., D. Garrick, W. Wu, M. Kearns & D. Martin, 1995. Position dependent variegation of globin transgene expression in mice. Proc. Natl. Acad. Sci. USA 92: 5371–5375.

Rossi, F.M., A.M. Krinstein, A. Spicher, O.M. Guicherit & H.M. Blau, 2000. Transcriptional control: rheostat converted to on/off switch. Mol. Cell 6: 723–728.

Roth, G., J. Blanke & D.B. Wake, 1994. Cell size predicts morphological complexity in the brains of frogs and salamanders. Proc. Natl. Acad. Sci. USA 91: 4796–4800.

Roth, G., K.C. Nishikawa & D.B. Wake, 1997. Genome size, secondary simplification, and the evolution of the brain in salamanders. Brain Behav. Evol. 50: 50–59.

Russanova, V.R., C.T. Driscoll & B.H. Howard, 1995. Adenovirus type 2 preferentially stimulates polymerase III transcription of *Alu* elements by relieving repression: a potential role for chromatin. Mol. Cell. Biol. 15: 4282–4290.

Sapienza, C., 1995. Genome imprinting: an overview. Dev. Genet. 17(3): 185–187.

Schmid, C.W., 1996. Alu: structure, origin, evolution, significance and function of one-tenth of human DNA. Prog. Nucl. Acid Res. Mol. Biol. 53: 283–319.

Schmid, C.W., 1998. Does SINE evolution preclude Alu function? Nucl. Acids Res. 26: 4541–4550.

Schmid, C.W. & K.C.J. Chen, 1985. The evolution of interspersed repetitive DNA sequences in mammals and other vertebrates, pp. 323–358 in Molecular Evolutionary Genetics, edited by R.J. MacIntire. Plenum Press, NY.

Schmid, C. & R. Maraia, 1992. Transcriptional regulation and transpositional selection of active SINE sequences. Curr. Opin. Genet. Dev. 2(6): 874–882.

Schmidt, T., 1999. LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. Plant Mol. Biol. 40: 903–910.

Schübeler, D., C. Francastel, D.M. Cimbora, A. Reik, D.I.K. Martin & M. Groudine, 2000. Nuclear localization and histone acetylation: a pathway for chromatin opening and transcriptional activation of the human β-globin locus. Genes Dev. 14: 940–950.

Schul, W., L. de Jong & R. Driel, 1998. Nuclear neighbors: the spatial and functional orgnization of genes and nuclear domains. J. Cell Biochem. 70: 159–171.

Scott, M. & A. Weiner, 1984. Structural relationships among genes that control development: Sequence homology between *Antennapedia*, *Ultrabithorax*, and *fushi tarazu* loci of *Drosophila*. Proc. Natl. Acad. Sci. USA 81: 4115–4119.

Seum, C., M. Delattre, A. Spierer & P. Spierer, 2001. Ectopic HP1 promotes chromosome loops and variegated silencing in *Drosophila.* EMBO J. 20: 812–818.

Shabalina, S.A. & A.S. Kondrashov. 1999. Pattern of selective constraint in *C. elegans* and *C. briggsae.* Genet. Res. Camb. 74: 23–30.

Shabalina, S.A., A.Y. Ogurtsov, V.A. Kondrashov & A.S. Kondrashov, 2001. Selective constraint in intergenic regions of human and mouse genomes. Trends Genet. 17: 373–376.

Sigrist, C.J. & V. Pirrotta, 1997. Chromatin insulator elements block the silencing of a target gene by the *Drosophila* polycomb response element (PRE) but allow *trans* interactions between PREs on different chromosomes. Genetics 147: 209–221.

Singh, P., 1994. Molecular mechanisms of cellular determination: their relation to chromatin structure and parental imprinting. J. Cell Sci. 107: 2653–2668.

Smit, A.F., 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. 9: 657–663.

Smit, A.F., G. Toth, A.D. Riggs & J. Jurka, 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J. Mol. Biol. 246: 401–417.

Smith, M.W., 1988. Structure of vertebrate genes: a statistical analysis implicating selection. J. Mol. Evol. 27: 45–55.

Spencer, H.G., A.G. Clark & M.W. Feldman, 1999. Genetic conflicts and the evolutionary origin of genomic imprinting. Tree 14: 197–201.

Spofford, J.B., 1976. Position-effect variegation in *Drosophila*, pp. 955–1018 in The Genetics and Biology of *Drosophila*, edited by M. Ashburner & E. Novitski. New York Academic Press, New York.

Stoltzfus, A., D.F. Spencer, M. Zuker, J.M. Logsdon Jr. & F.W. Doolittle, 1994. Testing the exon theory of genes: the evidence from protein structure. Science 268: 1366–1367.

Strutt, H., G. Cavalli & R. Paro, 1997. Co-localization of polycomb protein and GAGA factor on regulatory elements responsible

for the maintenance of homeotic gene expression. EMBO J. 16: 3621–3632.

Strutt, H. & R. Paro, 1997. The polycomb group protein complex of *Drosophila melanogaster* has different compositions at different target genes. Mol. Cell. Biol. 17: 6773–6783.

Stubblefield, E., 1986. A theory for developmental control by a program encoded in the genome. J. Theor. Biol. 118: 129–143.

Sutherland, H.G., M. Kearns, H.D. Morgan, A.P. Headley, C. Morris, D.I. Martin & E. Whitelaw, 2000. Reactivation of heritably silenced gene expression in mice. Mamm. Genome 11: 347–355.

Tchenio, T., J.F. Casella & T. Heidmann, 2000. Members of the SRY family regulate the human LINE retrotransposons. Nucl. Acids Res. 28: 411–415.

Thatcher J.W., J.M. Shaw & W.J. Dickinson, 1998. Marginal fitness contributions of nonessential genes in yeast. Proc. Natl. Acad. Sci. USA 95: 253–257.

Trelogan, S.A. & S.L. Martin, 1995. Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis. Proc. Natl. Acad. Sci. USA 92: 1520–1524.

Trifonov, E.N., 1989. The multiple codes of nucleotide sequences. Bull. Math. Biology 51: 417–432.

Trifonov, E.N., 1996. Interfering contexts of regulatory sequence elements. Cabios 12: 423–429.

Trimborn, T., J. Gribnau, F. Grosveld & P. Fraser, 1999. Mechanisms of developmental control of transcription in the murine α- and β-globin loci. Genes Dev. 13: 112–124.

Wallrath, L.L., 2000. *Drosophila* telomeric transgenes provide insights on mechanisms of gene silencing. Genetica 109: 25–33.

Walters, M.C., W. Magis, S. Fiering, J. Eidemiller, D. Scalzo, M. Groudine & D.I.K. Martin, 1996. Transcriptional enhancers act in *cis* to suppress position-effect variegation. Genes Dev. 10: 185–195.

Weiler, K.S. & Wakimoto, B.T., 1995. Heterochromatin and gene expression in Drosophila. Annu. Rev. Genet. 29: 577–605.

Weiner, A.M., 2000. Do all SINEs lead to LINEs? Nat. Genet. 24: 332–333.

Wessler, S.R., 1996. Turned on by stress. Plant retrotransposons. Curr. Biol. 6: 959–961.

Willard, H.F., 2001. Neocentromeres and human artificial chromosomes: an unnatural act. Proc. Natl. Acad. Sci. USA 98: 5374–5376.

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

Winzeler E.A. etc., and R.W. Davis, 1999. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science 285: 901–906.

Wolf, Y.I., F.A. Kondrashov & E.V. Koonin, 2000. No footprints of primordial introns in a eukaryotic genome. Trends Genet. 16: 333–334.

Xiong, Y. & T.H. Eickbush, 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. 9: 3353–3362.

Zink, D. & R. Paro, 1995. *Drosophila* polycomb-group regulated chromatin inhibits the accessibility of a *trans*-activator to its target DNA. EMBO J. 14: 5660–5671.

Zinkowski, R.P., J. Meyne & B.R. Brinkley, 1991. The centromere-kinetochore complex: a repeat subunit model. J. Cell. Biol. 113: 1091–1110.

Zuckerkandl E., 1974. A possible role of 'inert' heterochromatin in cell differentiation. Action of and competition for 'locking' molecules. Biochimie 56: 937–954.

Zuckerkandl, E., 1975. The appearance of new structures and functions in proteins during evolution. J. Mol. Evol. 7: 1–57.

Zuckerkandl, E., 1976. Evolutionary processes and evolutionary noise at the molecular level. I. J. Mol. Evol. 7: 167–183.

Zuckerkandl, E., 1978. Multilocus enzymes, gene regulation, and genetic sufficiency. J. Mol. Evol. 12: 57–89.

Zuckerkandl, E., 1981. A general function of noncoding polynuleotide sequences. Mass binding of transconformational proteins. Mol. Biol. Rep. 7: 149–158.

Zuckerkandl, E., 1983. Topological and quantitative relationships in evolving genomes, pp. 395–412 in Structure, Dynamics, Interactions and Evolution of Biological Macromolecules, edited by C. Hélène. D. Reidel Publishing Co.

Zuckerkandl, E., 1986. Polite DNA: functional density and functional compatibility in genomes. J. Mol. Evol. 24: 12–27.

Zuckerkandl, E., 1990. Can large insertions and deletions between genes affect development? J. Mol. Evol. 31: 161–162.

Zuckerkandl, E., 1991. Dispensability of parts of histones and the molecular clock. J. Mol. Evol. 32: 271–273.

Zuckerkandl, E., 1994. Molecular pathways of parallel evolution: I. Gene nexuses and their morphological correlates. J. Mol. Evol. 39: 661–678.

Zuckerkandl, E., 1997. Junk DNA and sectorial gene repression. Gene 205: 323–343.

Zuckerkandl, E., 1999. Sectorial gene repression in the control of development. Gene 238: 263–276.

Zuckerkandl, E., 2001. Intrinsically driven changes in gene interaction complexity. 1. Growth of regulatory complexes and increase in number of genes. J. Mol. Evol. 53: 539–554.

Zuckerkandl, E. & L. Pauling, 1965. Evolutionary divergence and convergence un proteins, pp. 97–165 in Evolving Genes and Proteins, edited by V. Bryson & H.J. Vogel. Academic Press, NY.

Zuckerkandl, E., G. Latter & J. Jurka, 1989. Maintenance of function without selection: *Alu* sequences as cheap genes. J. Mol. Evol. 29: 504–512.

Zuckerkandl, E. & W. Hennig, 1995. Tracking heterochromatin. Chromosoma 104: 75–83.