Biol. Rev. (2005), **80**, pp. 227–250. © 2005 *Cambridge Philosophical Society*
doi:10.1017/S1464793104006657 *Printed in the United Kingdom*

Why repetitive DNA is essential to genome function

James A. Shapiro[1,*] and Richard von Sternberg[2,3]

[1] *Department of Biochemistry and Molecular Biology*, *University of Chicago*, 920 *E.* 58*th Street*, *Chicago*, *IL* 60637, *USA* (E-mail: jsha@uchicago.edu)
[2] *National Center for Biotechnology Information – GenBank Building* 45, *Room* 6*AN*.18*D*-30, *National Institutes of Health*, *Bethesda*, *Maryland* 20894 (E-mail: sternber@ncbi.nlm.nih.gov)
[3] *Department of Systematic Biology*, *NHB*-163, *National Museum of Natural History*, *Smithsonian Institution*, *Washington*, *D.C.*, 20013-7012 (E-mail: Sternberg.Richard@NMNH.SI.EDU)

(*Received* 9 *March* 2004; *revised* 24 *September* 2004; *accepted* 29 *September* 2004)

ABSTRACT

There are clear theoretical reasons and many well-documented examples which show that repetitive DNA is essential for genome function. Generic repeated signals in the DNA are necessary to format expression of unique coding sequence files and to organise additional functions essential for genome replication and accurate transmission to progeny cells. Repetitive DNA sequence elements are also fundamental to the cooperative molecular interactions forming nucleoprotein complexes. Here, we review the surprising abundance of repetitive DNA in many genomes, describe its structural diversity, and discuss dozens of cases where the functional importance of repetitive elements has been studied in molecular detail. In particular, the fact that repeat elements serve either as initiators or boundaries for heterochromatin domains and provide a significant fraction of scaffolding/matrix attachment regions (S/MARs) suggests that the repetitive component of the genome plays a major architectonic role in higher order physical structuring. Employing an information science model, the 'functionalist' perspective on repetitive DNA leads to new ways of thinking about the systemic organisation of cellular genomes and provides several novel possibilities involving repeat elements in evolutionarily significant genome reorganisation. These ideas may facilitate the interpretation of comparisons between sequenced genomes, where the repetitive DNA component is often greater than the coding sequence component.

Key words: transposable element, non-coding DNA, satellite DNA, junk DNA, transcriptional regulation, chromatin domains, evolution, biocomputing, systems biology, data storage.

CONTENTS

* Address for Correspondence: Tel: 773-702-1625. Fax: 773-947-9345.

I. INTRODUCTION

(1) The conceptual problem posed by repetitive DNA

Fifty years of DNA-based molecular genetics and genome sequencing have revolutionised our ideas about the physical basis of cell and organismal heredity. We now understand many processes of genome expression and transmission in considerable molecular detail, and whole genome sequences allow us to think about the principles that underlie the organisation of cellular DNA molecules. There have been many surprises and new insights. In the human genome, for example, the protein-coding component represents about 1.2% of the total DNA, while 43% of the sequenced euchromatic portion of the genome consists of repeated and mobile DNA elements (International Human Genome Consortium, 2001; Table 1). In addition to dispersed elements, most of the unsequenced heterochromatic portion of the human genome (about 18% of the total) consists of repetitive DNA, both mobile elements and tandemly repeated 'satellite' DNA. Thus, over half the human genome is repetitive DNA. Table 1 shows that the human genome is far from exceptional in containing a major fraction of repeats. Even in bacteria, repetitive sequences may account for upwards of 5–10% of the total genome (Hofnung & Shapiro, 1999; Parkhill *et al.*, 2000).

Despite its abundance, the repetitive component of the genome is often called 'junk,' 'selfish,' or 'parasitic' DNA (Doolittle & Sapienza, 1980; Orgel & Crick, 1980). Because the view of repetitive and mobile genetic elements as genomic parasites continues to be influential (e.g. Lynch & Conery, 2003), we feel it is timely to present an alternative 'functionalist' point of view. The discovery of repetitive DNA presents a conceptual problem for traditional gene-based notions of hereditary information. This issue was noted in the pioneering work of Britten and colleagues (Britten & Kohne, 1968; Britten & Davidson, 1969, 1971; Davidson & Britten, 1979). We argue here that a more fruitful interpretation of sequence data may result from thinking about genomes as information storage systems with parallels to electronic information storage systems. From this informatics perspective, repetitive DNA is an essential component of genomes; it is required for formatting coding information so that it can be accurately expressed and for formatting DNA molecules for transmission to new generations of cells. In addition, the cooperative nature of protein-DNA interactions provides another fundamental reason why repeated sequence elements are essential to format genomic DNA. Instead of parasites, we argue that repetitive DNA elements are necessary organisers of genomic information.

(2) Overlooked aspects of genome organisation

Wide divergences between closely related taxa in repeat abundance and genome size (C-value) are often cited as evidence of the parasitic nature of repetitive DNA (Pagel & Johnstone, 1992; Vinogradov, 2003). This argument ignores three important aspects of genome organisation as a complex information system:

(a) The first aspect is that robust complex systems rely upon redundant components, many of which can be removed without detectably affecting overall system performance. Such robustness characterises repetitive DNA elements, particularly those arrayed in long tandem repeats that form compact nuclear structures, like centromeres. A minimum number may be required for function, but excesses are compatible with normal operation. Tandem arrays are often the regions that vary most between related taxa. Sometimes, different repeat elements fulfill equivalent tasks, so that absence of one particular element does not disable genome function.

(b) The second overlooked aspect is the significance of genome size and of distance between distinct regions of the genome. Rapidly reproducing organisms, like *Caenorhabditis*, *Drosophila*, *Fugu* and *Arabidopsis* tend to have stripped-down genomes with relatively less abundant repetitive DNA, while organisms with longer life cycles, such as humans and maize, have larger genomes with correspondingly more repetitive elements (Table 1; Cavalier-Smith, 1985).

(c) The third neglected aspect of genome organisation is the importance of stoichiometric balance between DNA-binding proteins and their cognate recognition elements (Schotta *et al.*, 2003). Changes in the ratio of proteins to repetitive elements influences chromatin structure and have observable phenotypic effects. Thus, repetitive DNA abundance is flexible but not adaptively neutral.

II. THE GENOME AS THE CELL'S LONG-TERM DATA STORAGE ORGANELLE: THE INFORMATICS METAPHOR

The key idea is to think of DNA as a sophisticated data storage medium. In order for cells to access, preserve, duplicate and transmit digitally encoded sequence information, DNA has to interact with other molecular components. Structurally and through its interactions with other

Table 1. DNA content in higher eukaryotes

Species	Genome size[1]	% repetitive DNA	% coding sequences	Reference
Animals				
Caenorhabditis elegans	100 MB	16.5	14	Stein *et al.* (2003)
Caenorhabditis briggsae	104 MB	22.4	13	Stein *et al.* (2003)
Drosophila melanogaster	175 MB	33.7 (female) ~57 (male)[2]	<10	Bennett *et al.* (2003); Celniker *et al.* (2002)
Ciona intestinalis	157 MB	35	9.5	Dehal *et al.* (2002)
Fugu rubripes	365 MB	15	9.5	Aparicio *et al.* (2002)
Canis domesticus	2.4 GB	31	1.45	Kirkness *et al.* (2003)
Mus musculus	2.5 GB	40	1.4	Mouse Genome Sequencing Consortium (2002)
Homo sapiens	2.9 GB	⩾50	1.2	International Human Genome Consortium (2001)
Plants				
Arabidopsis thaliana	125–157 MB	13–14	21	*Arabidopsis* Genome Initiative (2000) Bennett *et al.* (2003)
Oryza sativa (indica)	466 MB	42	11.8	Yu *et al.* (2002)
Oryza sativa (Japonica)	420 MB	45	11.9	Goff *et al.* (2002)
Zea mays	2.5 GB	77	1	Meyers *et al.* (2001)

[1] MB = megabases (10^6 base pairs), GB = gigabases (10^9 base pairs).
[2] The *D. melanogaster* Y chromosome is largely heterochromatic repetitive DNA.

cellular molecules, DNA stores information on three different time scales:

(1) **Long-term** ('genetic') storage involves DNA sequence information stable for many organismal generations.

(2) **Intermediate-term** ('epigenetic') storage occurs through complexing of DNA with protein and RNA into chromatin structures that may propagate over several cell generations (see e.g. Jenuwein & Allis, 2001, and other articles in the same issue; Van Speybroeck, Van de Vijver & De Waele, 2002). Chemical modifications of DNA that do not change sequence data, such as methylation, contribute to epigenetic storage (Bird, 2002).

(3) **Short-term** ('computational') information storage involves dynamic interactions of DNA with proteins and RNA molecules that can adapt rapidly within the cell cycle as the cellular environment changes.

The capacity of DNA to complex with RNA and protein to store information at these three different time scales enables the genome to fulfill multiple roles in cell and organismal heredity. Genomes serve as the organism's evolutionary record and most basic repository of specific phenotypic information. Genomes also participate in executing programs of cellular differentiation and multicellular morphogenesis during organismal life cycles. These programs are not hard-wired in the DNA sequence, and they sometimes permit the formation of very different organisms utilising a single genome (e.g. invertebrates having distinct larval and adult stages). Finally, genomes participate in computational responses that allow each cell to complete its cell cycle and deal with changing circumstances, correct internal errors and repair damage (Nurse, Masui & Hartwell, 1998).

The explicit parallel with electronic data systems indicates that the genomic storage medium has to be marked, or formatted, with generic signals so that operational hardware can locate and process the stored information. This is basically the idea of Britten and Davidson (1969). Like all information storage systems, the genome contains various classes of data files, and these files have to be formatted for access and copying. While many of the data files are unique (e.g. protein coding sequences), the formatting information must be far simpler in content. The requirement for reduced information content in formatting signals is the most basic reason that repetitive DNA sequences are essential to genome function.

III. MULTIPLE FUNCTIONS OF THE GENOME

It is tempting to reduce genome function to encoding cellular proteins and RNAs. However, molecular genetics has shown that achieving this task requires cells to possess a number of additional capabilities also encoded in the genome: (1) Regulating timing and extent of coding sequence expression. (2) Organizing coordinated expression of protein and RNA molecules that function together. (3) Packaging DNA appropriately within the cell. (4) Replicating the genome in synchrony with the cell division cycle. (5) Transmitting replicated DNA accurately to progeny cells at cell division. (6) Detecting and repairing errors and damage to the genome. (7) Restructuring the genome when necessary (as part of the normal life cycle or in response to a critical selective challenge). These additional capabilities involve specific kinds of interactions between DNA and other cellular molecules. The construction of highly precise transcription complexes in RNA and protein synthesis is one example of such interactions (Ptashne, 1986). Formation of a kinetochore structure at the centromere for attachment of microtubulues to ensure chromosome distribution at mitosis is another example (Volpe *et al.*, 2003).

Ever since the elaboration in the early 1960s of the operon model (Jacob & Monod, 1961) and the replicon hypothesis (Jacob, Brenner & Cuzin, 1963), we have understood specific interactions of cellular proteins with the genome to depend upon the existence of recognition signals in the DNA distinct from coding sequences. Signals like the operator or the origin of chromosome replication are completely different from any classical definition of a 'gene' as a basic unit. These recognition signals are themselves repetitive sequences in the genome, often carried by larger repeat elements. The idea that repetitive DNA is 'junk' without functional significance in the genome is simply not consistent with an extensive and growing literature, only a minor part of which is cited here.

IV. REPETITIVE SIGNALS IN THE HIERARCHICAL CONTROL OF CODING SEQUENCE EXPRESSION

To see how generic repetitive sequences format DNA for function, we can examine coding sequence expression. Historically, this genome operation has received the greatest attention. There are multiple roles for repetitive sequences. Specific transcription is not possible without generic start sites (promoters), stop sites (terminators in prokaryotes, polyA addition signals in eukaryotes) and signals for RNA processing (such as splice sites) (Alberts *et al.*, 2002). Regulation of transcription initiation involves arrays of binding sites for transcription factors that interact with the basic transcriptional apparatus to ensure proper timing and location of expression (Ptashne, 1986). When these binding sites are shared by several genetic loci, cells achieve coordinated expression of corresponding RNA and protein products (Arnone & Davidson, 1997).

Microbial genomes have an operon-like organisation at every scalar level (Audit & Ouzonis, 2003). This suggests that prokaryotic chromosomes are partitioned into a nested series of 'folders,' which can be differentially accessed by the cell. Likewise, eukaryotic genomes have a hierarchical structure that reflects at least three regulatory layers (van Driel, Fransz & Verschure, 2003). The existence of genomic folders has been described as epigenetic 'indexing' of the genome (Jenuwein, 2002). By regulation of chromatin remodeling and localisation of chromatin domains, cells achieve coordinate control over multi-locus segments of the genome, as observed in *Drosophila melanogaster* (Boutanaev *et al.*, 2002), *Caenorhabditis elegans* (Roy *et al.*, 2002), and humans (Caron *et al.*, 2001). In addition, systematic RNA inactivation (RNAi) knockout studies in *C. elegans* have identified extensive clusters of functionally related (hence, coordinately controlled) genetic loci (Kamath *et al.*, 2003). The molecular details of coordinate regulation by chromatin organisation are rapidly becoming evident (e.g. Greil *et al.*, 2003).

In eukaryotes, transcription is strongly influenced by how DNA is packaged. Packaging involves winding the DNA into nucleosomes, compacting the nucleosomes into higher order 'chromatin' structures complexed with protein and

Fig. 1. Linear structure of the *lac* operon regulatory region. The region schematised extends from the end of the *lacI* coding sequence to the beginning of the *lacZ* coding sequence. O1, O2 and O3 are the three documented operators; note that O2 is part of the *lacZ* coding sequence. The −10 and −35 regions of the canonical sigma 70 promoter are indicated together with the binding site for the cAMP receptor protein (CRP) needed for full promoter function.

RNA (Jenuwein & Allis, 2001), and attaching the chromatin to the nuclear scaffold or matrix (Laemmli *et al.*, 1992). Generic repetitive signals affect positioning of nucleosomes, formation of different classes of chromatin, extent of chromatin domains encompassing multiple genetic loci, and scaffold attachment (see Table 3).

Without entering into details here, it is widely recognised that the processes of replication, transmission, repair and DNA restructuring involve analogous recognition of specific repeated signals in the genome and the formation of higher-order nucleoprotein structures (Alberts *et al.*, 2002).

V. COOPERATIVE INTERACTIONS: A FUNDAMENTAL REASON FOR REPETITION IN GENOME FORMATTING

One of the simplest genome–proteome interaction systems, *E. coli lac* operon repression, provides a clear example of a fundamental principle: cooperative protein-DNA interactions involve repeated DNA sequence elements. Like many bacterial operons, *lac* contains multiple protein interactions sites, including related repeats of a sequence recognised by the LacI repressor's DNA-binding domain (Müller, Oehler & Müller-Hill, 1996; Fig. 1). As illustrated for the principal operator, O1, and for the distributed cyclic AMP receptor protein (CRP) binding site, these repeats are frequently organised as imperfect head-to-head palindromes (Ptashne, 1986; Fig. 2). The palindromic structure means that each recognition element is itself composed of repeats.

Operon repression requires that four repressor monomers organise into two dimers to bind four DNA recognition sequences in two palindromic operators (Lewis *et al.*, 1996; Fig. 2). One dimer binds each operator, and the two dimers contact each other through their C terminal domains. This binding forms a DNA loop which prevents RNA polymerase from accessing the promoter to initiate transcription. A single repressor monomer-half operator interaction is too weak to form a stable structure, but the cooperative interaction of two monomer-DNA binding events plus the protein–protein binding within the dimer creates a stable protein–DNA complex (Ptashne, 1986). Formation of the loop further stabilises the structure and more effectively occludes the promoter (Fig. 2). Thus, repeated copies of the recognition

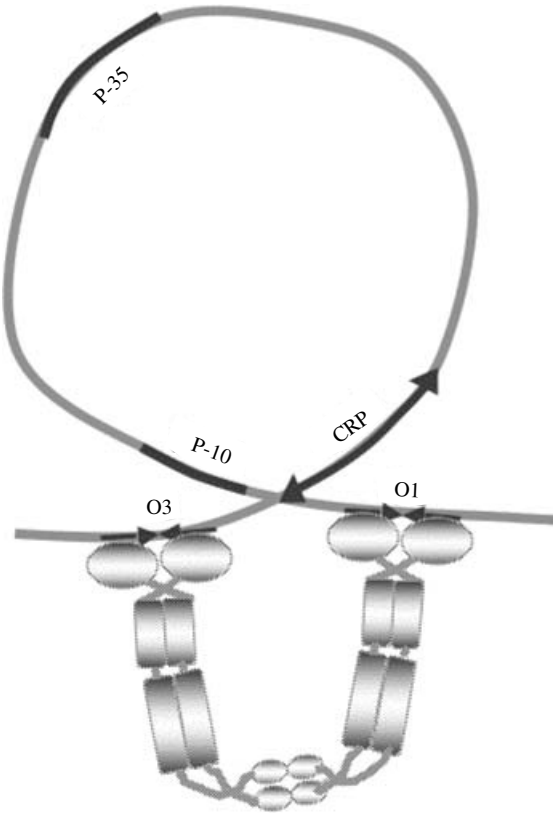

Fig. 2. Cartoon of *lac* promoter looping by repressor binding to O1 and O3 operators. The drawing is not to scale. Note that the effective promoter is divided into three separate components: the −10 and −35 binding sites for Sigma 70 RNA polymerase as well as the binding site for the required cAMP receptor protein (CRP) transcription factor.

sequence are critical to the formation of a repression complex.

When it is necessary to lift repression, an inducer molecule disrupts the cooperativity that supports a stable protein-DNA complex, repressor molecules separate from operators, and the promoter becomes accessible for transcription. Transcription initiation itself requires cooperative interactions; RNA polymerase binds to the −10 and −35 sites in the promoter and also to the dimeric cAMP-CRP complex, itself bound to two recognition sequences that make up the palindromic CRP site (Fig. 1).

The iteration of protein recognition sequences (transcription factor binding sites) is a general feature of both prokaryotic and eukaryotic transcriptional regulatory regions (Ptashne, 1986). In animals, regulatory regions contain multiple copies of several different binding sites (Arnone & Davidson, 1997). The relative arrangement of recognition sites determines whether protein binding will affect transcription positively or negatively. Different arrangements of several repeated recognition sites provides the combinatorial potential to construct a virtually infinite range of regulatory regions able to engage in sophisticated molecular computations with the cognate binding proteins, as documented in sea urchins and *Drosophila melanogaster* (Arnone & Davidson, 1997; Yuh, Bolouri & Davidson, 1998).

Cooperativity between individually weak but stereospecific molecular interactions to build meta-stable and stable multimolecular complexes or structures underlies the operation of cellular control circuits. Cooperativity and multiple interactions provide precision, combinatorial richness and robustness similar to the operation of multilayered fuzzy logic systems (http://zadeh.cs.berkeley.edu/). Because the underlying interactions are weak, multimolecular structures can assemble and disassemble rapidly to provide the dynamic responses needed to deal with a changing cellular environment. On the other hand, complexes with large numbers of cooperatively interacting components can be very stable, as occurs in longer-term epigenetic regulation of the genome. In the latter case, we find extensive arrays of repeated DNA elements in stable heterochromatic structures (Grewal & Elgin, 2002).

VI. STRUCTURAL VARIETIES OF REPETITIVE DNA

Repetitive DNA sequence elements involved in transcriptional regulation are chiefly oligonucleotide motifs recognised by specific DNA binding proteins. It is not essential that every copy be identical. Many tolerate differences in one or more base pairs and still provide binding specificity with an altered affinity for the cognate protein. Variations in binding sites serve regulatory functions, as in the classic example of the phage lambda operator (Ptashne, 1986). Since interaction of a protein factor with a recognition sequence motif is generally weak, the real functional entities in transcriptional formatting are composite elements that comprise two or more motifs, like promoters (Gralla & Collado-Vides, 1996), palindromic operators (Fig. 1), and enhancers (Arnone & Davidson, 1997).

In addition to oligonucleotide motifs and short composite elements (generally <100 base pairs), repetitive DNA sequences come in a wide variety of structural arrangements. Table 2 lists the most commonly recognised structures. Except for homopolymeric tracts and individual oligonucleotide motifs, repetitive elements are generally built up from simpler components. This modularity exemplifies the combinatorial and hierarchical nature of genomic coding resulting from cooperative interactions. Tandem array satellites are composed of sequence elements of defined length that typically contain motifs for binding of specific proteins, such as those required for chromatin configurations (Henikoff, Ahmad & Malik, 2001; Grewal & Elgin, 2002; Dawe, 2003). Tandem repetition creates extended structures where a very large number of cooperative interactions build up robust epigenetic stability persisting through multiple cell cycles.

Modularity also applies to transposons and retrotransposons that act to create novel long-term inherited DNA structures. Transposable elements contain signals that define the boundaries of each element and help create nucleoprotein structures that allow them to interact with

Table 2. Different structural classes of repetitive DNA

Structural class	Mnemonic	Structural or functional characteristics[1]
Oligonucleotide motif		4–50 bp; protein binding or recognition sites
Homopolymeric tract		Repeats of a single nucleotide $(N)_n$
VNTR	Variable nucleotide tandem repeats	Repeats of dinucleotides and longer sequences <100 bp that may vary in number in the tandem array; $(NN\ldots N)_n$
Composite elements		Composed of two or more oligonucleotide motifs, sometimes with non-specific spacer sequences; examples include palindromic operators, promoters, enhancers and silencers, replication origins, site-specific recombination sequences (Gralla & Collado-Vides, 1996; Craig *et al.*, 2002)
Tandem array satellites		Repeats of larger elements, typically 100–200 bp in length; satellite arrays typically contain thousands of copies (Henikoff *et al.*, 2001)
TIR DNA transposons	Terminal inverted repeat	DNA-based mobile genetic elements flanked by inverted terminal repeat sequences of $\leqslant 50$ bp; may encode proteins needed for transposition; vary in length from several hundred to several thousand base pairs (Craig *et al.*, 2002)
FB DNA transposons	Foldback	DNA transposons with extensive (many kb) inverted repeats at each end (Lim & Simmons, 1994)
Rolling circle DNA transposons		DNA transposons that insert from a circular intermediate by rolling circle replication; can generate tandem arrays (Kapitonov & Jurka, 2001; Craig *et al.*, 2002)
LTR retrotransposons	Long terminal repeat	Retroviruses and non-viral mobile elements flanked by direct terminal repeats of several hundred base pairs; insert at new locations following reverse transcription from an RNA copy into duplex DNA (Craig *et al.*, 2002)
LINE retrotransposons	Long interspersed nucleotide element	Mobile elements several kb in length with no terminal repeats; encode proteins involved in retrotransposition from a PolII-transcribed RNA copy by target-primed reverse transcription (Craig *et al.*, 2002)
SINE retrotransposons	Short interspersed nucleotide element	Mobile elements, a few hundred base pairs in length with no terminal repeats; do not encode proteins (mobilised by LINE products from a PolIII-transcribed RNA copy) (Craig *et al.*, 2002)

[1] bp = base pairs, kb = kilobase pairs.

and rearrange target DNA sequences [e.g. terminal inverted repeats (TIRs) of DNA transposons and long terminal repeats (LTRs) in retrotransposons (Craig *et al.*, 2002)]. In addition, transposons and retrotransposons contain sequence components that control transcription and may participate in DNA replication and chromatin organisation. We know from an extensive literature on insertional mutagenesis in nature and the laboratory that introduction of a transposable element into a particular location confers new functional properties on that region of the genome (Shapiro, 1983; Craig *et al.*, 2002; Deininger *et al.*, 2003).

VII. DOCUMENTATION OF DIVERSE GENOMIC FUNCTIONS ASSOCIATED WITH DIFFERENT CLASSES OF REPETITIVE DNA ELEMENTS

Table 3 provides a compilation of repetitive DNA functions in genome operation. These range from basic transcription, through regulation at transcriptional and post-transcriptional levels, to chromatin and nuclear organisation, genome transmission at cell division, damage repair and DNA restructuring. In some cases, functional categories overlap. For example, DNA restructuring at homopolymeric tracts and variable tandem nucleotide repeats (VNTRs) in pathogenic bacteria and chromatin organisation in eukaryotes both serve as mechanisms for regulating coding sequence expression. Moreover, recombination mechanisms fulfill roles in repair [double-strand (DS) break correction] as well as cellular protein engineering [antigenic variation, V(D)J joining (Bassing, Swat & Alt, 2002), and immunoglobulin class switching]. Particularly noteworthy are the cases where repetitive DNA influences the physical organisation and movement of the genome through the cell cycle. These cases include the delineation of centromeres by tandem repeat arrays, the formation of telomeres by non-LTR retrotransposons, delineation of chromatin domains by boundary/insulator elements, and the attachment of DNA to the nuclear matrix by sequences found in larger composite elements, such as mammalian long interspersed nucleotide elements (LINEs).

Table 3. Selected examples of repetitive DNA functions

Function	Structural class[1]	Example	Comment	Reference
TRANSCRIPTION				
Promoters	Transposable elements (TE)		TE sequences in almost a quarter of human promoter sequences	Jordan *et al.* (2003)
	LINE	Human LINE-1	1.6 % of 2004 examined human promoters include LINEs; the 5′-untranslated region of L1 has both an internal (sense) promoter and an antisense promoter (ASP); L1 ASP chimeric transcripts are highly represented in expressed-sequence tag (EST) databases	Speek (2001); Zaiss & Kloetzel (1999); Nigumann *et al.* (2002); Jordan *et al.* (2003)
	SINE	Human *Alu*s; mouse *B2* elements	Genomic synonyms; RNA polymerase II promoter elements – 5.3 % of 2004 examined human promoters have SINEs as components	Ferrigno *et al.* (2001); Jordan *et al.* (2003)
	Unclassified middle repetitive	RENT elements (repetitive element from Nicotiana tabacum)	>5 kb with conserved 5′ ends but variable 3′ termini; moderately repetitive (~100 copies); found only in certain *Nicotiana* species	Foster *et al.* (2003)
	Unclassified	CpG islands	Characteristic clusters of dinucleotides associated with mammalian promoters	Ioshikhes & Zhang (2000); Ponger & Mouchiroud (2002)
Enhancers & Silencers	LINE	Human Line-1	Positive transcriptional regulatory element; binding sites for testis-determining transcription factors	Yang *et al.* (1998); Becker *et al.* (1993); Tchénio *et al.* (2000)
	SINE	Subgroup II–III of human *Alu*Sx subfamily	Nuclear hormone receptor binding sites for thyroid hormone receptor, retinoic acid receptor and estrogen receptor	Norris *et al.* (1995); Vansant & Reynolds (1995); Babich *et al.* (1999)
		Jo, Jb, Sq, Sp, Sx, and Sg subfamilies of human *Alu*s; subset of rodent *B1* elements	Genomic synonyms; Pax6 binding sites	Zhou *et al.* (2000, 2002)
	VNTR	*Drosophila* (GA)n and (GAGA)n elements	*Bithoraxoid* polycomb group response elements are (GA)n repeats; ~250 *Drosophila* loci have GAGA elements; the *Drosophila* GAGA factor (GAF) binds to the 5′ and intronic regions of loci with (GA)n sequences	Hodgson *et al.* (2001); Mahmoudi *et al.* (2002); van Steensel *et al.* (2003)
		Mammalian (GAGA)n elements	Found in RNA PolII and PolIII promoter elements in *Alu* elements; enriched binding sites for various mammalian chromatin/transcriptional factors; negative regulatory effects mediated by microsatellites and VNTRs can range from general to cell-type-specific	Humphrey *et al.* (1996); Kropotov *et al.* (1997, 1999); Tomilin *et al.* (1992); Tsuchiya *et al.* (1998); Cox *et al.* (1998); Albanese *et al.* (2001); Fabregat *et al.* (2001); Rothenburg *et al.* (2001); Youn *et al.* (2002)
		Mammalian triplet repeats	Repeat unit number differences result in quantitative variation in gene silencing/down-regulation in many instances (VNTRs as expression 'tuning forks,'; Kashi *et al.* 1997)	Saveliev *et al.* (2003); Tovar *et al.* (2003)
		Arabidopsis and other plant (GA)n elements		Santi *et al.* (2003); Sangwan & O'Brian (2002)

Table 3 (*cont.*)

Function	Structural class[1]	Example	Comment	Reference
	Unclassified	*S. cerevisiae* subtelomeric X elements	Core X [470 nt] can enhance the action of a distant silencer without acting as a silencer on its own	Lebrun *et al.* (2001)
	Unclassified	*Drosophila* subtelomeric satellite-like repeat TAS	A complex subterminal satellite with a 457-bp repeat unit (also called telomere-associated sequence, TAS) silences adjacent sequences	Kurenova *et al.* (1998); Boivin *et al.* (2003)
	Oligonucleotide motif	*Schizosaccharomyces pombe* Cre (cAMP-response-element) = ATGACGT and related sequences	Almost 1000 Cre hotspots are dispersed throughout the *S. pombe* genome. Meiotic recombination hotspot activity	Fox *et al.* (2000)
	Unclassified	Unnamed *Arabidopsis thaliana* repetitive elements	Act as enhancers	Ott & Hansen (1996)
	Unclassified	*Strongylocentrotus* RSR elements	Act as enhancers	Gan *et al.* (1990)
Transcription attenuation	Mosaic repetitive elements	*E. coli* BIMEs (bacterial interspersed mosaic elements)	BIMEs are part of a Rho-dependent transcriptional regulational system involving up to 250 operons	Espeli *et al.* (2001)
	LINE	L1	Retards transcript elongation	Han *et al.* (2004)
Terminators	Oligonucleotide motif	AATAAA	Canonical polyadenylation signal	Wahle & Ruegsegger (1999); Barabino & Keller (1999)
	TIR DNA transposons	IS elements	Rho-dependent and Rho-independent terminators	http://www-is.biotoul.fr/is.html
Regulatory RNAs	LTR retrotransposon	Mouse VL30 elements	Non-protein coding transcripts of VL30 elements selectively bind to PSF (pre-mRNA splicing factor) repressor, allowing transcription of genes controlled by insulin-like growth factor response elements (IGFRE); the VL30 transcripts are causally involved in steroidogenesis and oncogenesis	Song *et al.* (2004)
POST-TRANSCRIPTIONAL RNA PROCESSING				
mRNA targeting	SINE	Rodent *ID* elements	Target mRNAs to neuronal dendrites; genomic synonym	Chen *et al.* (2003)
		Rodent *BC200* and primate homologue	Neuronal targeting; genomic synonym	Skryabin *et al.* (1998)
		Primate *Alu*	Neuronal targeting; genomic synonym	Watson & Sutcliffe (1987)
	Unclassified	*Xenopus laevis* Xlslirts familiy	(3–13 dispersed copies of a 79–81 bp monomer unit); transcripts localise RNAs to the vegetal cortex	Kloc & Etkin (1994); Zearfoss *et al.* (2003)
TRANSLATION				
Selective enhancement of mRNA translation	SINE	Human *Alu*s; mouse *B1*, *B2* elements	Genomic synonyms	Rubin *et al.* (2002)
DNA REPLICATION, LOCALISATION AND MOVEMENT				
Repication origins (Ori)	Oligonucleotide motifs	Bacterial chromosomes and plasmids	Multiple nearby repeats in origins, specific to each replicon	del Solar *et al.* (1998); Marczynski & Shapiro (1993)

		A3/4 = CCTCAAATGGTC TCCATTTTCCTTT GGCAAATTCC	ORS (origin recognition sequence)	Schild-Poulter *et al.* (2003); Novac *et al.* (2001); Price *et al.* (2003)
	LTR and unclassified	*S. cerevisiae* LTR, subtelomeric X and Y′ repeats	Contain 20 % of *S. cerevisiae* sequences that immunoprecipitate with origin recognition proteins	Wyrick *et al.* (2001)
Centromeres	Tandem array satellites		Large tandem arrays form pericentric heterochromatin, facilitate centromere function	Choo (2001); Henikoff *et al.* (2001)
		4–5 kb dg-dh repeats in *S. pombe*	RNAi, dg-dh DS RNA needed for centromere organisation; capable of inducing silencing at an ectopic site	Jenuwein (2002); Dawe (2003); Volpe *et al.* (2002); Hall *et al.* (2002); Reinhart & Bartel (2002)
		171 bp alpha repeats in primates	Alpha-satellite arrays are highly competent human artifical chromosome (HAC)-forming substrates	Grimes *et al.* (2002); Schueler *et al.* (2001); Vafa & Sullivan (1997)
		180 bp *Arabidopsis* repeat		Nagaki *et al.* (2003*b*); Copenhaver *et al.* (1999)
		156 bp CentC in maize	CENH3 replaces histone H3 in centromeres, and CentC interacts specifically with CENH3	Ananiev *et al.* (1998*a*); Zhong *et al.* (2002); Nagaki *et al.* (2003*a*)
		155 bp CentO rice repeat		Cheng *et al.* (2002)
	LTR	Cereal centromere repeats		Aragon-Alcaide *et al.* (1996)
		CRR	Centromere-specific retrotransposon in rice	Cheng *et al.* (2002)
		Maize CRM; CentA, Huck and Prem2	CENH3 interacts specifically with CRM (centromere retrotransposon in maize)	Ananiev *et al.* (1998*b*); Zhong *et al.* (2002); Nagaki *et al.* (2003*a*)
		Arabidopsis Athila element		Pelissier *et al.* (1996)
		250, 301 bp repeats in wheat and rye	Ty3/*gypsy*-related	Cheng & Murata (2003)
		Ty3/*gypsy* family *Sorghum* elements	Ty3/*gypsy*-related sequences present exclusively in the centromeres of all *Sorghum* chromosomes; Ty1/*copia*-related DNA sequences are not specific to the centromeric regions	Miller *et al.* (1998)
Meiotic pairing and recombination	VNTR	(CAGG)10-18, (CAGA)4-6 in mice	Meiotic recombinational hotspots in mouse MHC (major histocompatability) region	Isobe *et al.* (2002); Shiroishi *et al.* (1995)
	Unclassified	*Drosophila* heterochromatin	Heterochromatin regions initiate pairing in male meiosis; depends upon rDNA spacer repeats	McKee *et al.* (2000)
	Oligonucleotide motif	*Schizosaccharomyces pombe* Cre (cAMP-response-element) = ATGACGT and related sequences	Almost 1000 Cre hotspots are dispersed throughout the *S. pombe* genome. Meiotic recombination hotspot activity	Fox *et al.* (2000)
Telomeres	non-LTR retrotransposons	*Drosophila* heterochromatin telomere repeat A (HeT-A), telomere-associated retrotransposon (TART)	HeT-A units work in pairs: the 5′ element has a promoter in the 3′ untranscribed region that allows transcription of the adjacent template-unit	Pardue & DeBaryshe (2003)
		Plasmodium telomere associated repetitive elements (TAREs)	*Plasmodium falciparum* telomere-associated sequences of the 14 linear chromosomes display a similar higher order organisation and form clusters of four to seven telomeres localised at the nuclear periphery	Figueiredo *et al.* (2000, 2002)

Table 3 (*cont.*)

Function	Structural class[1]	Example	Comment	Reference
		Giardia telomere retrotransposons		Arkhipova & Morrison (2001)
	176, 340, or 350 bp DNA repeats	*Chironomus* spp.	Maintain telomere length by recombination/gene conversion mechanisms; repeat lengths species-specific	Cohn & Edstrom (1992)
	LTR retrotransposon	Yeast Ty1	Ty1 activated when normal telomere function impaired	Scholes *et al.* (2003)
S/MARs (scaffold/ matrix associated regions)	TIR DNA Transposons	Rice and sorghum genome miniature inverted repeat transposable elements (MITEs)	Most of the MARs discovered in the two genomic regions co-localise with MITEs	Avramova *et al.* (1998)
		Transposable *Euplotes crassus* (*Tec*) elements	S/MARs that undergo en masse, developmental, chromosomal elimination	Sharp *et al.* (2003)
	LTR	Human LTR retrotransposons	7.0 % of examined human S/MARs derived from LTR retrotransposons	Jordan *et al.* (2003)
		Drosophila gypsy	Elements determining intranuclear gene localisation/ nuclear pore association	Gerasimova *et al.* (2000); Labrador & Corces (2002)
	LINEs	Human LINE-1	39.4 % of human S/MARs are LINE sequences; 98 LINE1 consensus sequences were found to contain 14 distinct S/MAR recognition signatures; the distribution of *Alu* and LINE repetitive DNA are biased to positions at or adjacent to apoptotic cleavage sites; LINE1 elements retard transcript elongation	Chimera & Musich (1985); Rollini *et al.* (1999); Khodarev *et al.* (2000); Jordan *et al.* (2003); Han *et al.* (2004)
CHROMATIN ORGANISATION, NUCLEAR ARCHITECTURE & EPIGENETIC MODIFICATION				
Nucleosome positioning elements	VNTR	(TATAAACGCC)n	Flexible DNA segments, highest documented affinity for nucleosomes, cluster around centromeres; the TATA tetrad (5′-TATAAACGCC-3′), is found in multiple phased repeats in genomic sequences that are among the strongest known binders of core histones	Widlund *et al.* (1999)
Heterochromatin	DNA Transposons; LTR and non-LTR retroposons	*Drosophila* transposable elements	Nine transposable elements (*copia*, *gypsy*, *mdg*-1, *blood*, *Doc*, *I*, *F*, *G*, and *Bari*-1) are preferentially clustered into one or more discrete heterochromatic regions in chromosomes of the Oregon-R laboratory stock; *P* and *hobo* elements, recent invaders of the *D. melanogaster* genome exhibit heterochromatic clusters in certain natural populations	Cryderman *et al.* (1998); Pimpinelli *et al.* (1995)
	LTR retroposon	Hamster intracisternal A particle (IAP) elements	In Syrian hamster, over half of the genomic IAP elements are accumulated in heterochromatin, including along the entire Y chromosome	Dimitri & Junakovic (1999)
		Maize *Grande*, *Prem2*, *RE-10*, *RE-15*, and *Zeon*	Abundant in heterochromatic knob regions; blocks of tandem 180-bp repeats interrupted by insertions of full size copies of retrotransposable elements; about 30 % of cloned knob DNA fragments	Ananiev *et al.* (1998*c*)
		Arabidopsis Athila	*Athila* elements in the *Arabidopsis* genome are concentrated in or near heterochromatic regions. Most of the heterochromatic elements retrotransposed directly into 180 bp satellite clusters	Pelissier *et al.* (1996)

		Several *Drosophila* LTR families	LTR elements represent 61 % of euchromatic transposable elements and approximately 78 % of heterochromatic elements. LINE elements represent 24 % of the euchromatic and 17 % of the heterochromatic transposable element sequence. DNA elements represent 15 % in euchromatin and 5 % in heterochromatin	Hoskins *et al.* (2002)
	LINE	Human LINE-1	X inactivation, monoallelic expression, imprinting	Lyon (2000); Parish *et al.* (2002); Allen *et al.* (2003)
Methylation	Oligonucleotide motif	GATC	DNA adenine methylase (DAM) site in Gram-negative bacteria; involved in methylation control of transcription, replication, repair, chromosome packaging and transposition	Barras & Marinus (1989); Low *et al.* (2001)
	Tandem repeat satellite	180 bp, 350 bp TR1 repeats in maize knobs		Ananiev *et al.* (1998*a*)
	SINE	Mouse *B1*	*B1* elements methylated de novo to a high level after transfection into embryonal carcinoma cells; *B1* elements acted synergistically	Yates *et al.* (1999)
	Unclassified – palindromic repeat	Petunia repetitive sequence (RPS) element	The palindromic RPS element acts as a de novo hypermethylation site in the non-repetitive genomic background of *Arabidopsis*	Müller *et al.* (2002)
	Unclassified	CpG islands	Methylation sites clustered near eukaryotic promoters; methylation indicates silenced state	Ioshikes & Zhang (2000); Robertson (2002)
Insulator/Boundary elements	Unclassified	*Saccharomyces* subtelomeric anti-silencing repeats (STARs)	The telomere-proximal portion of either X or Y′ dampened silencing when located between the telomere and the reporter gene and also at one of the silenced mating-type cassettes	Fourel *et al.* (1999); Pryde & Louis (1999)
	LTR	*Drosophila gypsy* element	The *gypsy* insulator blocks propagation of silencing and alters the nuclear localisation of adjacent DNA	Gerasimova *et al.* (2000); Byrd & Corces (2003)
	Unclassified	*Drosophila* boundary element 28 (BE28)	Approximately 150 copies in the *D. melanogaster* genome; this 269 bp BE is part of a 1.2 kb repeated sequence. BE28 element maps to several pericentric chromosomal regions, blocks promoter-enhancer interactions in a directional manner, and binds the boundary-element associated DNA-binding factor BEAF	Cuvier *et al.* (2002)
ERROR CORRECTION AND REPAIR				
Homolgous recombination (double-strand break repair)	Oligonucleotide motif	*Chi* sites in *E. coli*, *B. subtilis*, *H. influenzae* and *Lactococcus lactis*	Distinct recombination initiation signals in each bacterial species	El-Karoui *et al.* (1999); Chedin *et al.* (2000); Quiberoni *et al.* (2001)
Methyl-directed mismatch repair (MMR)	Oligonucleotide motif	GATC	Dam methylation site, binds MutH mutational repair protein when hemimethylated	Modrich (1989)

Table 3 (*cont.*)

Function	Structural class[1]	Example	Comment	Reference
DNA RESTRUCTURING				
Antigenic variation	Oligonucleotide motifs	*M. bovis vis* (35 bp)	Inversions generate multiple distinct variant surface proteins by site-specific recombination	Lysnyansky, Ron & Yogev (2001)
	Unclassified	*N. gonorrhoeae* and *N. meningitidis* NIME (*Neisseria* interspersed mosaic element) repeats flanking silent pilus (pil) cassettes	Recombination involving flanking repeats replaces segments of pilus coding sequence at expressed *pil* locus	Parkhill *et al.* (2000); Saunders *et al.* (2000)
		Borrelia downstream homology sequence (DHS) 200 bp and 17–18 bp repeats	Surface protein variation by expression site switching	Wang *et al.* (1995); Barbour *et al.* (2000)
Phase variation	Homopolymeric tracts	*Neisseria meningitidis opc* locus, *N. gonorrhoeae pilC*	Phase variation of opacity (*opc*) and pilus (*pilC*) proteins	Sarkari *et al.* (1994); Jonsson *et al.* 1991; van der Ende *et al.* (1995); Henderson *et al.* (1999); Bayliss, Field & Moxon (2001)
	VNTR	*N. meningitidis hmbR* locus	Phase variation of outer membrane haemoglobin-binding protein (Hmb)	Richardson & Stojiljkovic (1999)
		H. influenzae adhesins	Tandem heptamers (ATCTTTC)	Dawid *et al.* (1999)
		N. gonorrhoeae Opa proteins	Tandem pentamers (CTCTT)	Stern & Meyer (1987)
Global genome plasticity	VNTR	Various *Helicobacter pylori* repeats	Repetitive, nonrandomly positioned VNTRs act as recombination centers promoting host-specific genomic modification	Aras *et al.* (2003*a*, *b*)
Uptake and integration of laterally transferred DNA	Oligonucleotide motif	*V. cholerae* repeats (VCRs)	Construction of pathogenicity operons by site-specific recombination	Mazel *et al.* (1998)
		DNA uptake: *N. gonorrhoeae* = GGCGTCTGAA; *H. influenzae* and *Actinobacillus actinomycetemcomitans* = AAGTGCGGTCA	Sequences identifying conspecific DNA	Chen & Dubnau (2003)
Chromatin diminution	DNA transposons	Excised elements in *Paramecium* and some *Oxytrichia* species; *Euplotes crassus* Tec elements	DNA transposon-like elements that undergo en masse, developmental, chromosomal elimination	DuBois & Prescott (1997); Prescott (2000); Sharp *et al.* (2003)
VDJ recombination	Oligonucleotide motif composites	RAG (recombination-associated gene) transposase recognition sequences (RSSs)	Protein engineering, rapid protein evolution. Recombination (DNA DS break) signals in mammalian immune systems; composed of 7 bp – 12/23 bp spacer – 9 bp elements	Bassing, Swat & Alt (2002); Gellert (2002)
Immunoglobulin class switching	VNTR	S (switch) regions upstream of immunoglobulin heavy chain constant region exons	Protein engineering. Tandem arrays that undergo DS breaks when transcribed from lymphokine-controlled promoters	Kitao *et al.* (2000); Kinoshita & Honjo (2001)

[1] See Table 2 for the definitions of abbreviations/mnemonics for the various repeat structural classes.

One feature of Table 3 is that certain repetitive elements, such as the mammalian LINE-1 and *Drosophila gypsy* retrovirus/retrotransposon (Pelisson *et al.*, 2002), appear under multiple functional headings. For example, the most intensively studied feature of *gypsy* is the role it plays in the organisation of chromatin domains. It contains an 'insulator' or boundary element capable of separating inactive and active chromatin domains by tethering the DNA to the nuclear matrix, thereby creating a discontinuity in the chromatin structure (Gerasimova, Byrd & Corces, 2000; Byrd & Corces, 2003). Since *gypsy* and other mobile elements retain their structures as they migrate through the genome, there is predictability to the signals they will carry with them. Thus, cells have the ability to introduce a pre-organised constellation of functional signals into any location in the genome. Although *gypsy* is often rare in *Drosophila* genomes and may be dispensible, the *D. melanogaster* genome also has approximately 150 boundary elements at the base of each chromosome containing the 269 bp BE28 sequence inside an unclassified 1.2 kb repeat element. Apparently, therefore, repetitive elements play a significant role in the physical organisation of *Drosophila* chromatin.

The LINE-1 element in the human genome has externally oriented promoter and polyadenylation activity, scaffold/matrix attachment region (S/MAR) signals, and has been implicated in X chromosome inactivation by facilitating heterochromatin formation. Very recently, a LINE1 role in modulating transcription has been discovered (Han, Szak & Boeke, 2004). Since about 850 000 LINE elements make up a remarkable 21 % of human euchromatic DNA, it is difficult to deny that the LINE elements constitute major architectonic and expression-related features of our hereditary material. The notion that LINE elements are major organisers of genome functional architecture is supported by close comparative analysis of syntenic regions in the mouse and human genomes. In two aligned segments, 18/25 and 9/11 murine L1 elements have putative human orthologues in the same orientation (Zhu, Swergold & Seldin, 2003). The high degree of conservation in the positions and orientations of highly variable elements implies positive functional selection.

Even the very stripped-down, repeat-poor genome of the yeast *S. cerevisiae* has boundary elements embedded in the subtelomeric Y′ repeats. Intriguingly, these boundary elements form part of a larger complex that also includes silencing elements in the adjacent X repeats and an autonomously replicating sequence (ARS) that serves as a site for the initiation of DNA replication. About 20 % of putative yeast ARS sequences correspond to LTR regions of complete or defective retrotransposons (Wyrick *et al.*, 2001). Thus, it appears that repetitive elements play an important role in organizing the chromatin and replication structures of the budding yeast genome.

VIII. SYNTENY AND GENOMIC SYNONYMS

Comparative whole-genome sequencing is beginning to provide additional supporting evidence for the structural/organisational roles of repetitive elements by documenting the existence of extensive 'synteny' between related genomes as well as the presence of 'genomic synonyms' at comparable locations in related genomes. Two 'genomic synonyms' are evolutionarily independent (i.e. unrelated) elements belonging to the same structural class that play the same functional role in each species. 'Synteny' refers to the occurrence of large genome segments, often megabases in length, that share the same order of genetic loci and are assumed to represent evolutionary conservation of a functional genomic arrangement (Eichler & Sankoff, 2003). The locations of related LINE and short interspersed nucleotide element (SINE) sequences are also conserved in syntenic regions of the mouse and human genomes (Silva *et al.*, 2003), indicating positive selection.

Widespread examples of genomic synonyms are the 100–200 bp elements composing long tandem repeat arrays in pericentromeric heterochromatin that delineate centromeres. More narrowly defined genomic synonyms are *Alu* and *B1* SINE elements of primate and rodent genomes. These elements derive from partial dimers of the 7S RNA sequence, but they arose independently or diverged from a common structure before either was amplified in the primate and rodent genomes. *Alu* is not found in rodents, and B1 is not found in primates (Schmid, 1996; Vassetzky, Ten & Kramerov, 2003). Both *Alu* and *B1* elements have been documented to carry similar transcriptional and translational regulatory signals. *Alu* elements are also genomic synonyms to mouse *B2* elements as promoters, and *Alu*, *B1* and *B2* serve as synonyms in enhancing mRNA translation. Primate *Alu* further serves as a synonym for rodent 'identifier' (*ID*) and *BC200* elements targeting specific transcripts in nerve cells (see Table 3 for individual references).

Because the many thousands of copies of primate *Alu*s and rodent *B1*s, *B2*s, *ID* and *BC200s* are structurally distinct, their distributions in the human and mouse genomes originated in each species from literally thousands of distinct retrotransposition events at some point after the divergence of the rodent and primate ancestors. Despite independent amplifications, the coarse-grained distributions of SINE elements are similar within syntenic regions (Fig. 12 in Mouse Genome Sequencing Consortium, 2002), supporting the proposition that locations of these synonyms have related functions in both genomes. Within the human genome, functional roles have been invoked to account for the nonrandom localisation of *Alu* in genetic loci involved in metabolism, signaling, and transport (Grover *et al.*, 2003) and for observed nonrandom higher-order patterns (Versteeg *et al.*, 2003).

A significant test of the genomic synonym hypothesis will be to see what patterns, if any, emerge from the results of a fine-grained comparison between the locations of *Alu* in the human genome and its putative synonyms, *B1*, *B2*, *ID* and *BC200* in the mouse genome.

IX. REPEATS AND RNAi IN HETEROCHROMATIN FORMATTING

The role of repetitive elements as genomic synonyms in heterochromatin formatting poses an apparent paradox.

How do conserved proteins involved in establishing heterochromatin recognise distinct DNA sequences? Part of the answer appears to be that initial repeat recognition is accomplished by means of cognate RNAs rather than by proteins (Jenuwein, 2002). It has long been known that fungi can detect genomic repetitions independently of sequence content and specifically methylate repeated sequences (Selker, 1990). Similar methylated repeats were discovered in transgenic plants and connected mechanistically to the presence of complementary RNA (Hamilton & Baulcombe, 1999; Mette *et al.*, 2000). DNA methylation is a step in the formation of transcriptionally inactive heterochromatin from transcriptionally active euchromatin (Bird, 2002). These observations were some of the first indications of RNA-directed inactivation, or RNAi (Matzke, Matzke & Kooter, 2001).

The role of RNAi in repeat-directed heterochromatin formation has been confirmed by genetic studies in the fission yeast *Schizosaccharomyces pombe*. Mutants in the *dcr*, *RdRp*, and *ago* loci defective in the RNAi RNA-processing and silencing machinery display aberrations in heterochromatin formation and accumulate double-stranded transcripts from the tandem array repeats flanking the centromeres (Volpe *et al.*, 2002). In wild-type *S. pombe* cells, a majority of short RNA fragments characteristic of RNAi correspond to pericentromeric repeat sequences (Reinhart & Bartel, 2002). Thus, it appeared there might be a mechanistic link between RNAi and chromatin formatting by repetitive sequences.

Taking advantage of the ability to manipulate the yeast genome, the RNAi – repeat-formatting hypothesis was tested by inserting a reporter locus into pericentromeric heterochromatin arrays. In wild-type cells, the inserted reporter is not expressed, but expression is derepressed in the *dcr*, *RdRp*, and *ago* mutants (Volpe *et al.*, 2002). Thus, RNAi is required to establish inactive heterochromatin around centromeres. In a complementary experiment, a centromere-related *CenH* element was inserted in a normally euchromatic region. Ectopically, *CenH* silenced a linked reporter and induced histone modifications characteristic of heterochromatin (Hall *et al.*, 2002). Introduction of *dcr*, *RdRp*, and *ago* mutations eliminated the silencing. Thus, RNA recognition of repeats can lead to heterochromatin formatting in a manner that has yet to be determined. RNAi is necessary for centromere function in *S. pombe* (Volpe *et al.*, 2003). A broad range of observations in plant and animal systems indicate that the RNAi–repetitive DNA-heterochromatin connection is widespread among eukaryotes (summarised in Dawe, 2003; Martienssen, 2003, and in Verdel *et al.*, 2004), as recently confirmed in *Drosophila melanogaster* (Pal-Bhadra *et al.*, 2004).

X. TAXONOMICALLY RESTRICTED GENOME SYSTEM ARCHITECTURE

The view of the genome advocated here as a hierarchically organised data storage system formatted by repetitive DNA sequence elements implies that each organism has a genome system architecture, in the same way that each computer data storage system has a characteristic architecture. In the computer example, architecture depends upon the operating system and hardware that are used, not upon the content of each data file. Macintosh©, Windows© and Unix© machines can all display the same images and text files, even though the data retrieval paths are operationally quite distinct. Similarly, many protein and RNA sequences (data files) are conserved through evolution, but different taxa organise and format their genomes in quite different ways for replication, transmission and expression. An overall system architecture is required since these processes must be coordinated so that they operate without mutual interference. DNA segments must be in the right place at the right time for function. Chromatin formatting for large-scale organisation of transcription and replication domains is well documented (Jenuwein & Allis, 2001, and other articles in the same issue; van Driel *et al.*, 2003), and we are learning about higher levels of spatio-temporal organisation of transcription and replication into 'factory' zones (Rui, 1999; Sawitzke & Austin, 2001; Bentley, 2002).

One clear example of taxonomically-specific genome system architectures involves the different signals used for basic transcription in bacteria and archaea, organisms sharing many coding sequences that are formatted completely differently for transcription. These include, notably, the heat-shock *hsp70* locus encoding a group of three orthologous chaperone proteins (Macario *et al.*, 1999). Another example of differential transcription formatting concerns catabolite repression. The generic signal marking catabolite-repressed sequences in *Escherichia coli* (the CRP palindromic binding site for the CRP-cAMP complex in Fig. 1) is completely different from its genomic synonym in *Bacillus subtilis* (CRE element recognised by the catabolite control protein CcpA; Miwa *et al.*, 2000). While both *E. coli* and *B. subtilis* use orthologous transport systems to monitor external glucose, independently evolved molecular signal transduction paths connect them to the catabolite repression signals in the genome.

An aspect of genome system architecture that deserves special discussion is distribution of dispersed and clustered repetitive DNA elements along the chromosomes. Table 3 shows how these elements are rich in signals affecting transcription, RNA processing, chromatin organisation, and attachment of the DNA to the nuclear matrix. We know that the genomic locations of these repeats have significant effects on function. There is a vast literature on multiple phenotypic changes caused by transposable element insertions, even outside well-defined genetic loci (Shapiro, 1983; Craig *et al.*, 2002), and by the phenomenon known as 'position effect variegation' (PEV) (Spofford, 1976; Henikoff, 1996; Wakimoto, 1998). PEV is observed when heterochromatic blocks of clustered repeats inhibit expression of adjacent genetic loci. PEV effects occur over megabase distances. PEV thus constitutes a clear demonstration that the repeat content of each genome is an important aspect of system architecture.

It is important to note here the little known fact that phenotypic effects of heterochromatin are not necessarily limited to adjacent genetic loci. The strength of heterochromatic silencing on the three large *D. melanogaster*

chromosomes is sensitive to the total nuclear content of heterochromatin carried on the Y chromosome (see Table 1). Increase in Y chromosome dosage (XYY males) suppresses PEV silencing, while reduction in the number of Y chromosome repeats (XO males) enhances PEV silencing (Spofford, 1976; Dimitri & Pisano, 1989). Changing dosage of repetitive DNA encoding ribosomal RNA in the nucleolar organiser (NO) region had similar effects (Spofford & DeSalle, 1991). We now understand these inter-chromosomal effects on developmental coding sequence expression to result from titration of heterochromatin-specific proteins in the nucleus (Henikoff, 1996; Schotta *et al.*, 2003). Consequently, the sizes of repeat sequence arrays are not neutral. Indeed, expansion and contraction of major heterochromatic blocks may serve as higher-level 'tuning forks' for developmental processes in the same way that VNTR expansion and contraction regulate single locus expression (Kashi, King & Doller, 1997; Trifonov, 1999). Comparable *trans* effects on aspects of genome functioning have been observed with B chromosomes in maize (Carlson, 1978).

The architectural role of dispersed repeats agrees with the conservation detected in the positions and orientations of shared repetitive elements (Silva *et al.*, 2003; Zhu *et al.*, 2003). The observations on conserved repeats suggest that high numbers of 'framework elements' may be retained in disparate mammalian genomes, with more derived subfamilies of LINEs, SINEs, and LTR elements being restricted to particular families and genera. Moreover, genome analysis is beginning to provide evidence of functional roles related to imprinting for evolutionarily 'recent' LINE element insertions. Nonorthologous LINE-1 elements are similarly positioned asymmetrically in the X-inactivation centres of human, mouse, and cow (Chureau *et al.*, 2002), and L1 elements are significantly associated with monoallelically expressed loci in both human and mouse genomes (Allen *et al.*, 2003).

From a perspective postulating that changes in repetitive elements may be important events in establishing specific new genome architectures, it is significant to note that repetitive DNA can be far more taxonomically discriminating than coding sequences. For example, each order of mammals has its own characteristic set of SINE elements (Weiner, Deininger & Efstratiadis, 1986; Sternberg & Shapiro, in press). Since these highly iterated SINEs are independently derived from cellular sequences, such as different tRNA or 7S RNA sequences, it is clear that taxonomic diversification among mammals involved many thousands of independent SINE amplification and insertion events. Similarly, plant species can be discriminated by their pericentromeric repeats (Table 3). The related nematode species *C. elegans* and *C. briggsae* have genomes composed, respectively, of 16.5 % and 22.4 % repeat DNA, but any one of the ten major repeat elements from either species is not found in the other (Stein *et al.*, 2003). Sibling species of *Drosophila* often share both morphology and protein polymorphisms, but they can still be can be identified because they contain different simple sequence satellite DNAs as well as different abundances of particular transposable elements (Dowsett, 1983; Bachmann & Sperlich, 1993; Miller *et al.*, 2000). Operationally, it is much easier to identify the species of origin of a DNA, cell culture, or tissue sample by examining repetitive DNA than coding or unique sequences, and this principle of using repeats for identification is applied within species for forensic DNA analysis (Jeffreys *et al.*, 1993).

Another frequently ignored feature of genome system architecture associated with repeat elements is overall genome size (Cavalier-Smith, 1985). In plants, genome size correlates with an increase in repetitive DNA abundance (Table 1). Plant molecular geneticists have suggested that the total length of each genome is an important functional characteristic, which influences replication time, a characteristic that correlates with the length of the life cycle (Bennett, 1998; Bennetzen, 2000; Petrov, 2001; Vinogradov, 2003). It makes sense that amplification of mobile genetic elements is an efficient method of altering total DNA content in the genome. Similarly, distance between regulatory and coding sequences may be an important control parameter (Zuckerandl, 2002).

XI. DISCUSSION

(1) Genome system architecture and evolution

The concept of genome system architectures formatted by repetitive DNA extends the range of conceivable changes that confer adaptive benefits or reproductive isolation. This idea obliges us to consider the effects of altering repetitive components of the genome as well as unique coding and regulatory sequences. Classical evolutionary theory assumes that phenotypic variation involves alteration of individual gene products due to changes in coding sequences. We now know this view is too restricted in two ways. First of all, protein structure can change without altering the coding sequences themselves through rearrangement of exons or *via* alteration of splicing patterns. Formation of new exon combinations by segmental duplications represents one class of such changes (Eichler, 2001), and insertion of repetitive elements into introns to alter splicing patterns is another (Nekrutenko & Li, 2001). Segmental duplications, insertion of repetitive elements, exon shuffling (Moran, DeBerardinis & Kazazian, 1999) as well as major chromosome rearrangements (Gray, 2000; Bailey, Liu & Ecihler, 2003) are all processes mediated by dispersed repetitive elements.

The second way that the classical view is unnecessarily restricted is in constraining adaptive variation to changes in an organism's repertoire of protein and RNA. Changes in regulatory formatting of conserved coding sequences can alter developmental patterns and lead to new traits using the same repertoire of proteins and RNAs (Britten & Davidson, 1971). The *Drosophila* literature is replete with examples of major morphological changes caused by alterations in repetitive elements.

Examination of genomic sequences indicates that rearrangement of repetitive elements has played a significant role in adaptive evolution and multicellular development. In Table 3 we noted the use of a VNTR element to format

ontogenetic silencing of the Polycomb response element (PRE) in the Ultrabithorax region (Hodgson, Argiropoulis & Brock, 2001). A search of the human genome reveals many examples of regulatory regions evolved from repetitive elements (Britten, 1996; Brosius, 1999; Jordan *et al.*, 2003). These conclusions from genome scanning agree with detailed molecular genetic analyses that demonstrate the participation of repetitive elements in regulation of coding sequence expression (e.g. Mozer & Benzer, 1994; Song, Sui & Garen, 2004). Moreover, as we have seen, changes in repetitive DNA affect chromatin formatting and nuclear organisation and thus have consequences for developmental expression of large chromosomal regions.

The genome system architecture perspective predicts a major role for evolutionary diversification by alterations in repetitive DNA that alter genome transmission without affecting phenotype. This is just what we find in groups of closely related and sibling species, which may display no detectable morphological, physiological or adaptive differences. In the *Cyclops* group of copepods, large heterochromatic blocks have different chromosomal locations in each species and undergo distinct patterns of excision during early somatic development (Beermann, 1977; Wyngaard & Gregory, 2001).

Changes in centromeric repeats constitute another situation where repetitive DNA variations may alter chromosome transmission but not organismal phenotypes. In some cases, such as the Indian Muntjac deer, we know that germ line incompatibility with the parental species resulted from genome restructuring. The *Muntiacus muntjak* deer ancestor underwent one or more Robertsonian fusions of centromeric heterochromatin that reduced chromosome numbers and cause abnormal pairing and non-disjunction at meiosis in hybrids with sibling SE Asian deer species (Fontana & Rubini, 1990; He & Brinkley, 1996).

Since major changes in control of genome maintenance and transmission can occur by altering repetitive sequences that format these processes, such alterations can lead to reproductive isolation and set the stage for subsequent phylogenetically-restricted changes in phenotype. In other words, we suggest that major evolutionary events can *initiate* within the repetitive sector of the genome. They do not have to follow changes in the coding sector. The importance of repetitive elements as major actors in evolutionary diversification has been expounded most forcefully by Dover's concept of 'molecular drive' (Dover, 1982). Importantly, however, we disagree with his viewing changes in repetitive DNA as functionally neutral. Indeed, we argue that functionality is precisely what makes repetitive elements powerful agents of taxonomic separation.

(2) A more integrative view of the genome

It is commonly accepted that the major information in genomes consists of coding sequences determining protein and RNA molecules. The importance of regulatory signals has also been widely recognised, and searches for phylogenetically conserved non-coding sequences constitute an active subfield of genomics. Nonetheless, the conceptual significance of these 'non-coding' components of the genome has largely gone unnoticed, with the result that acceptance of the 'selfish DNA' hypothesis is rarely challenged.

As we enter the era of 'systems biology' (Kitano, 2002), it is useful to recall that a system is more than a collection of components. Those components need to integrate functionally so they can accomplish systemic tasks requiring cooperative action. One way to state our argument is to say that repetitive DNA elements provide the physical basis within the genome for functional integration. As Britten and Davidson (1969, 1971) realized, dispersed regulatory sites connect unlinked coding sequences into coordinately controlled subsystems. Similarly, replication and genome transmission processes are organised by generic signals that determine origins, telomeres, centromeres and other nucleoprotein complexes involved in genome maintenance. Signals for formatting and delineating various chromatin domains provide a higher level of organisation for both transcription and replication, and distributed sites for attachment to cellular or nuclear structures provide a dynamic overall physical organisation of the genome that we are just beginning to comprehend.

A second consequence of an integrative view of repetitive genome components will be to focus attention on their importance in evolution. There has been growing recognition of the role of mobile elements as agents of DNA restructuring (summarised in Shapiro, 1999*a*, *b*, 2002*a*, *b*), but less consideration about questions of how repetitive elements come to be distributed as they are. Analysis of individual genomes indicates that there have been episodes of expansion by particular subfamilies of SINE elements (International Human Genome Consortium, 2001; Mouse Genome Sequencing Consortium, 2002). These episodes probably indicate periods of rapid evolutionary change and may help clarify the punctuated nature of the evolutionary record. It may also prove worthwhile to search for cases where changes in repetitive elements have been the major engine of speciation, as appears to be the case in copepods and Muntjac deer.

(3) Repetitive DNA and the computational metaphor for the genome

We report only a small portion of the growing literature on the functional roles of repetitive DNA elements. The trend is clearly towards discovering greater specificity, pattern and significance in the surprisingly abundant repeat fraction of genomes. As we increasingly apply computational metaphors to cellular function, we expect that a deeper understanding of repetitive elements, the integrative fraction of cellular DNA, will reveal novel aspects of the logical architecture inherent to genome organisation.

The electronic computation metaphor can only be applied so far to cellular information processing. The former is a digitized process based on binary coding and Turing machine principles (http://www.abelard.org/turpap/turpap.htm), while the latter is a poorly-understood analog process based on molecular stereospecificity and templating as well as on sequence encoding. In the case of repetitive DNA, we encounter limitations to the metaphor because

this genomic fraction has aspects of both software and hardware. Repetitive DNA acts like software insofar as it is encoded in the DNA sequence and is utilised by the cell many times to carry out defined routines, such as heterochromatin condensation. Like software, repetitive DNA can control operations involving different unique data files. On the other hand, repetitive DNA also forms part of essential cellular machinery, such as the mitotic apparatus. When DNA serves as a physical substrate facilitating protein aggregation during nucleoprotein complex formation in transcription, recombination and chromosome packaging, it operates as hardware.

The fact that simple distinctions between software and hardware do not readily apply to the informatics of repetitive DNA holds an important and encouraging lesson. In the era of biocomputing and systems biology, our study of cellular information processing promises to revolutionise not only the life sciences but also the information sciences. We can anticipate learning powerful new computational paradigms as we come to understand how cells use myriad molecular components to regulate millions of biochemical events that occur every minute of every cell cycle. Indeed, we may come one day to regard erstwhile 'junk DNA' as an integral part of cellular control regimes that can truly be called 'expert.'

XII. CONCLUSIONS

(1) DNA is a data storage medium, and genomes function as computational information organelles.

(2) Proper access to coding sequence data files and reliable genome replication and transmission to progeny cells require that there be generic repetitive signals to format the DNA for interaction with cellular hardware. Repeat signal formatting is also necessary for genome packaging, repair and restructuring.

(3) Generic repetitive signals are distinct from the genes envisioned by conventional theory and require us to employ informatic metaphors in conceptualizing fundamental principles of genome organization. Repetitive DNA elements provide the physical basis for integrating different regions of the genome and for coordinating interdependent aspects of genome function (e.g. packaging, expression and transmission).

(4) Cooperative interactions between repeated DNA elements and iterated protein domains are essential to the formation of nucleoprotein complexes that carry out basic genome operations. The cooperative nature of molecular interactions in cellular information processing provides a second fundamental reason why repetitive DNA elements are essential to genome function.

(5) There is extensive documentation in the molecular genetic literature (some of it tabulated here) that all structural varieties of repetitive DNA play significant roles in one or more categories of genomic tasks. More complex repetitive DNA elements, such as retrotransposons, carry integrated sets of signals influencing multiple functions. Evolutionarily independent elements can serve as 'genome synonyms' to provide similar functional formatting in different species.

(6) The distribution of repetitive signals confers a characteristic genome system architecture independent of coding sequence content. Different genome system architectures can have distinct transmission and expression properties even with the same coding sequences. Meaningful evolutionary change can take place in the repetitive component of the genome without altering coding sequences.

(7) Recognition by repeat-derived small interfering siRNA molecules provides a mechanistic basis for analogous chromatin formatting by repetitive arrays with different sequence contents.

XIII. ACKNOWLEDGMENTS

We thank Adam Wilkins, Michael Ashburner and an anonymous reviewer for helpful comments on the manuscript. J. A. S. wishes to thank members of NORDITA and the Niels Bohr Institute, Copenhagen, for inviting him to participate in summer schools that helped begin his interdisciplinary thinking about biological issues. R. V. S. wishes to thank Dr Michael J. Dewey, University of South Carolina, for bringing repetitive DNA elements to his attention so many years ago and for encouraging him to study these enigmatic sequences.

XIV. REFERENCES

Albanese, V., Biguet, N. F., Kiefer, H., Bayard, E., Mallet, J. & Meloni, R. (2001). Quantitative effects on gene silencing by allelic variation at a tetranucleotide microsatellite. *Human Molecular Genetics* **10**, 1785–1792.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. (2002). *The Molecular Biology of the Cell*. 4th Edition. Garland Scientific, Taylor & Francis, New York.

Allen, E., Horvath, S., Tong, F., Spiteri, E., Riggs, A. D. & Marahrens, Y. (2003). High concentrations of long interspersed nuclear element sequence distinguish monoallelically expressed genes. *Proceedings of the National Academy of Sciences, USA* **100**, 9940–9945.

Ananiev, E. V., Phillips, R. L. & Rines, H. W. (1998*a*). A knob-associated tandem repeat in maize capable of forming fold-back DNA segments: are chromosome knobs megatransposons? *Proceedings of the National Academy of Sciences, USA* **95**, 10785–10790.

Ananiev, E. V., Phillips, R. L. & Rines, H. W. (1998*b*). Chromosome-specific molecular organisation of maize (*Zea mays* L.) centromeric regions. *Proceedings of the National Academy of Sciences, USA* **95**, 13073–13078.

Ananiev, E. V., Phillips, R. L. & Rines, H. W. (1998*c*). Complex structure of knob DNA on maize chromosome 9: retrotransposon invasion into heterochromatin. *Genetics* **149**, 2025–2037.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., *et al.* (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815.

Aragon-Alcaide, L., Miller, T., Schwarzacher, T., Reader, S. & Moore, G. (1996). A cereal centromeric sequence. *Chromosoma* **105**, 261–268.

Aras, R. A., Fischer, W., Perez-Perez, G. I., Crosatti, M., Ando, T., Haas, R. & Blaser, M. J. (2003*a*). Plasticity of repetitive DNA sequences within a bacterial (Type IV) secretion system component. *Journal of Experimental Medicine* **198**, 1349–1360.

Aras, R. A., Kang, J., Tschumi, A. I., Harasaki, Y. & Blaser, M. J. (2003*b*). Extensive repetitive DNA facilitates prokaryotic genome plasticity. *Proceedings of the National Academy of Sciences, USA* **100**, 13579–13584.

Arkhipova, I. R. & Morrison, H. G. (2001). Three retrotransposon families in the genome of *Giardia lamblia*: two telomeric, one dead. *Proceedings of the National Academy of Sciences, USA* **98**, 14497–14502.

Arnone, M. I. & Davidson, E. H. (1997). The hardwiring of development: organisation and function of genomic regulatory systems. *Development* **124**, 1851–1864.

Audit, B. & Ouzonis, C. A. (2003). From genes to genomes: universal scale-invariant properties of microbial chromosome organisation. *Journal of Molecular Biology* **332**, 617–633.

Avramova, Z., Tikhonov, A., Chen, M. & Bennetzen, J. L. (1998). Matrix-attachment regions and structural colinearity in the genomes of two grass species. *Nucleic Acids Research* **26**, 761–767.

Babich, V., Aksenov, N., Alexeenko, V., Oei, S. L., Buchlow, G. & Tomilin, N. (1999). Association of some potential hormone response elements in human genes with Alu family repeats. *Gene* **239**, 341–349.

Bachmann, L. & Sperlich, D. (1993). Gradual evolution of a specific satellite DNA family in *Drosophila ambigua*, *D. tristis*, and *D. obscura*. *Molecular Biology and Evolution* **10**, 647–659.

Bailey, J. A., Liu, G. & Eichler, E. E. (2003). An Alu transposition model for the origin and expansion of human segmental duplications. *American Journal of Human Genetics* **73**, 823–834.

Barabino, S. M. L. & Keller, W. (1999). Last but Not Least: Regulated Poly(A) Tail Formation. *Cell* **99**, 9–11.

Barbour, A. G., Carter, C. J. & Sohaskey, C. D. (2000). Surface protein variation by expression site switching in the relapsing fever agent *Borrelia hermsii*. *Infection and Immunity* **68**, 7114–7121.

Barras, F. & Marinus, M. G. (1989). The great GATC: DNA methylation in *E. coli*. *Trends in Genetics* **5**, 139–143.

Bassing, C. H., Swat, W. & Alt, F. W. (2002). The mechanism and regulation of chromosomal V(D)J recombination. *Cell* **109**, S45–S55.

Bayliss, C. D., Field, D. & Moxon, E. R. (2001). The simple sequence contingency loci of *Haemophilus influenzae* and *Neisseria meningitidis*. *Journal of Clinical Investigation* **107**, 657–662.

Becker, K. G., Swergold, G. D., Ozata, K. & Thayer, R. E. (1993). Binding of the ubiquitous nuclear transcription factor YY1 to a *cis* regulatory sequence in the human LINE-1 transposable element. *Human Molecular Genetics* **2**, 1697–1702.

Beermann, S. (1977). The diminution of heterochromatic chromosomal segments in *Cyclops* (Crustacea, Copepoda). *Chromosoma* **60**, 297–344.

Bennett, M. D. (1998). Plant genome values: how much do we know? *Proceedings of the National Academy of Sciences, USA* **95**, 2011–2016.

Bennett, M. D., Leitch, I. J., Price, H. J. & Johnston, J. S. (2003). Comparisons with Caenorhabditis (approximately 100 Mb) and Drosophila (approximately 175 Mb) using flow cytometry show genome size in Arabidopsis to be approximately 157 Mb and thus approximately 25% larger than the Arabidopsis genome initiative estimate of approximately 125 Mb. *Annals of Botany* **91**, 547–557.

Bennetzen, J. L. (2000). Transposable elements contributions to plant gene and genome evolution. *Plant Molecular Biology* **42**, 251–269.

Bentley, D. (2002). The mRNA assembly line: transcription and processing machines in the same factory. *Current Opinion in Cell Biology* **14**, 336–342.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes and Development* **16**, 6–21.

Boivin, A., Gally, C., Netter, S., Anxolabehere, D. & Ronsseray, S. (2003). Telomeric associated sequences of *Drosophila* recruit polycomb-group proteins *in vivo* and can induce pairing-sensitive repression. *Genetics* **164**, 195–208.

Boutanaev, A. M., Kalmykova, A. I., Shevelyov, Y. Y. & Nurminsky, D. I. (2002). Large clusters of co-expressed genes in the *Drosophila* genome. *Nature* **420**, 666–669.

Britten, R. J. (1996). DNA sequence insertion and evolutionary variation in gene regulation. *Proceedings of the National Academy of Sciences, USA* **93**, 9374–9377.

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

Britten, R. J. & Davidson, E. H. (1971). Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Quarterly Review of Biology* **46**, 111–138.

Britten, R. J. & Kohne, D. E. (1968). Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* **161**, 529–540.

Brosius, J. (1999). RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* **238**, 115–134.

Byrd, K. & Corces, V. G. (2003). Visualization of chromatin domains created by the gypsy insulator of *Drosophila*. *Journal of Cell Biology* **162**, 565–574.

Carlson, W. R. (1978). The B chromosome of corn. *Annual Reviews of Genetics* **12**, 5–23.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M. C., van Asperen, R., Boon, K., Voute, P. A., Heisterkamp, S., van Kampen, A. & Versteeg, R. (2001). The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291**, 1289–1292.

Cavalier-Smith, T. (1985). *The evolution of genome size*. John Wiley & Sons Ltd., Chichester.

Celniker, S. E., Wheeler, D. A., Kronmiller, B., Carlson, J. W., Halpern, A., Patel, S., Adams, M., Champe, M., Dugan, S. P., Frise, E., *et al.* (2002). Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. Genome Biol. 2002;3(12):RESEARCH0079. Epub 2002 Dec 23.

Chedin, F., Ehrlich, S. D. & Kowalczykowski, S. C. (2000). The *Bacillus subtilis* AddAB Helicase/Nuclease is regulated by its cognate Chi sequence *in vitro*. *Journal of Molecular Biology* **298**, 7–20.

Chen, I. & Dubnau, D. (2003). DNA transport during transformation. *Frontiers in Bioscience* **8**, S544–S556.

Chen, D., Kunlin, J., Kawaguchi, K., Nakayama, M., Zhou, X., Xiong, Z., Zhou, A., Mao, X. O., Greenberg, D. A., Graham, S. H. & Simon, R. P. (2003). Ero1-L, an ischemia-inducible gene

from rat brain with homology to global ischemia-induced gene 11 (Giig11), is localized to neuronal dendrites by a dispersed identifier (ID) element-dependent mechanism. *Journal of Neurochemistry* **85**, 670–679.

Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, C. R., Gu, M., Blattner, F. R. & Jiang, J. (2002). Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell* **14**, 1691–1704.

Cheng, Z.-J. & Murata, M. (2003). A centromeric tandem repeat family originating from a part of Ty3/gypsy-retroelement in wheat and its relatives. *Genetics* **164**, 665–672.

Chimera, J. A. & Musich, P. R. (1985). The association of the interspersed repetitive KpnI sequences with the nuclear matrix. *Journal of Biological Chemistry* **260**, 9373–9379.

Choo, K. H. (2001). Domain organisation at the centromere and neocentromere. *Developmental Cell* **1**, 165–177.

Chureau, C., Prissette, M., Bourdet, A., Barbe, V., Cattolico, L., Jones, L., Eggen, A., Avner, P. & Duret, L. (2002). Comparative sequence analysis of the X-inactivation center region in mouse, human, and bovine. *Genome Research* **12**, 894–908.

Cohn, M. & Edstrom, J.-L. (1992). Telomere-associated repeats in Chironomus form discrete subfamilies generated by gene conversion. *Journal of Molecular Evolution* **35**, 114–122.

Copenhaver, G. P., Nickel, K., Kuromori, T., Benito, M. I., Kaul, S., Lin, X., Bevan, M., Murphy, G., Harris, B., Parnell, L. D., McCombie, W. R., Martienssen, R. A., Marra, M., Preuss, D., *et al.* (1999). Genetic definition and sequence analysis of Arabidopsis centromeres. *Science* **286**, 2468–2474.

Cox, G. S., Gutkin, D. W., Haas, M. J. & Cosgrove, D. E. (1998). Isolation of an Alu repetitive DNAbinding protein and effect of CpG methylation on binding to its recognition sequence. *Biochimica et Biophysica Acta* **1396**, 67–87.

Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M. (2002). *Mobile DNA II*. ASM Press, Washington DC.

Cryderman, D. E., Cuaycong, M. H., Elgin, S. C. R. & Wallrath, L. L. (1998). Characterization of sequences associated with position-effect variegation at pericentric sites in *Drosophila* heterochromatin. *Chromosoma* **107**, 277–285.

Cuvier, O., Hart, C. M., Käs, E. & Laemmli, U. K. (2002). Identification of a multicopy chromatin boundary element at the borders of silenced chromosomal domains. *Chromosoma* **110**, 519–531.

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

Dawe, R. K. (2003). RNA interference, transposons, and the centromere. *Plant Cell* **15**, 297–301.

Dawid, S., Barenkamp, S. J. & Stgeme, W. J. (1999). Variation in expression of the *Haemophilus influenzae* HMW adhesins: a prokaryotic system reminiscent of eukaryotes. *Proceedings of the National Academy of Sciences, USA* **96**, 1077–1082.

Dehal, P., Satou, Y., Campbell, R. K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D. M., *et al.* (2002). The draft genome sequence of *Ciona intestinalis*: insights into chordate and vertebrate orgins. *Science* **298**, 2157–2167.

Deininger, P. L., Moran, J. V., Batzer, M. A. & Kazazian, H. H. (2003). Mobile elements and mammalian genome evolution. *Current Opinion in Genetics and Development* **13**, 651–658.

Del Solar, G., Giraldo, R., Ruiz-Echevarría, M. J., Espinosa, M. & Díaz-Orejas, R. (1998). Replication and control of circular bacterial plasmids. *Microbiology and Molecular Biology Reviews* **62**, 434–464.

Dimitri, P. & Junakovic, N. (1999). Revising the selfish DNA hypothesis: new evidence on accumulation of transposable elements in heterochromatin. *Trends in Genetics* **15**, 123–124.

Dimitri, P. & Pisano, C. (1989). Position effect variegation in *Drosophila melanogaster*: relationship between suppression effect and the amount of Y chromosome. *Genetics* **122**, 793–800.

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

Dover, G. A. (1982). Molecular drive: a cohesive mode of species evolution. *Nature* **299**, 111–117.

Dowsett, A. P. (1983). Closely related species of *Drosophila* can contain different libraries of middle repetitive DNA sequences. *Chromosoma* **88**, 104–108.

Dubois, M. L. & Prescott, D. M. (1997). Volatility of internal eliminated segments in germ line genes of hypotrichous ciliates. *Molecular and Cellular Biology* **17**, 326–337.

Eichler, E. E. (2001). Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends in Genetics* **17**, 661–669.

Eichler, E. E. & Sankoff, D. (2003). Structural dynamics of eukaryotic chromosome evolution. *Science* **301**, 793–797.

El Karoui, M., Biaudet, V., Schbath, S. & Gruss, A. (1999). Characteristics of Chi distribution on different bacterial genomes. *Research in Microbiology* **150**, 579–587.

Espeli, O., Moulin, L. & Boccard, F. (2001). Transcription attenuation associated with bacterial repetitive extragenic BIME elements. *Journal of Molecular Biology* **314**, 375–386.

Fabregat, I., Koch, K. S., Aoki, T., Atkinson, A. E., Dang, H., Amosova, O., Fresco, J. R., Schildkraut, C. H. & Leffert, H. L. (2001). Functional pleiotropy of an intramolecular triplex-forming fragment from the 3′-UTR of the rat Pigr gene. *Physiological Genomics* **5**, 53–65.

Ferrigno, O., Virolle, T., Djabari, Z., Ortonne, J. P., White, R. J. & Aberdam, D. (2001). Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. *Nature Genetics* **28**, 77–81.

Figueiredo, L. M., Pirrit, L. A. & Scherf, A. (2000). Genomic organisation and chromatin structure of *Plasmodium falciparum* chromosome ends. *Molecular and Biochemical Parasitology* **106**, 169–174.

Figueiredo, L. M., Freitas-Junior, L. H., Bottius, E., Olivo-Martin, J.-C. & Schert, A. (2002). A central role for *Plasmodium falciparum* subtelomeric regions in spatial positioning and telomere length regulation. *EMBO Journal* **21**, 815–824.

Fontana, F. & Rubini, M. (1990). Chromosomal evolution in Cervidae. *Biosystems* **24**, 157–174.

Foster, E., Hattori, J., Zhang, P., Labbe, H., Martin-Heller, T., Li-Pook-Than, J., Ouellet, T., Mallet, K. & Miki, B. (2003). The new RENT family of repetitive elements in *Nicotiana* species harbors gene regulatory elements related to the tCUP cryptic promoter. *Genome* **46**, 146–155.

Fourel, G., Revardel, E., Koering, C. E. & Gilson, E. (1999). Cohabitation of insulators and silencing elements in yeast subtelomeric regions. *EMBO Journal* **18**, 2522–2537.

Fox, M. E., Yamada, T., Ohta, K. & Smith, G. R. (2000). A family of cAMP-response-element-related DNA sequences with meiotic recombination hotspot activity in *Schizosaccharomyces pombe*. *Genetics* **156**, 59–68.

Gan, L., Zhang, W. & Klein, W. H. (1990). Repetitive DNA sequences linked to the sea urchin spec genes contain

transcriptional enhancer-like elements. *Developmental Biology* **139**, 180–196.

Gellert, M. (2002). V(D)J recombination: RAG proteins, repair factors, and regulation. *Annual Review of Biochemistry* **71**, 101–132.

Gerasimova, T. I., Byrd, K. & Corces, V. G. (2000). A chromatin insulator determines the nuclear localizations of DNA. *Molecular Cell* **6**, 1025–1035.

Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., *et al.* (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**, 92–100.

Gralla, J. D. & Collado-Vides, J. (1996). Organisation and function of transcription regulatory elements. In *Escherichia coli and Salmonella Cellular and Molecular Biology*, 2nd ed., (eds. F. C. Neidhardt *et al.*), pp. 1232–1245. ASM Press, Washington, D.C.

Gray, Y. H. (2000). It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends in Genetics* **16**, 461–468.

Greil, F., Van Der Kraan, I., Delrow, J., Smothers, J. F., De Wit, E., Bussemaker, H. J., Van Driel, R., Henikoff, S. & Van Steensel, B. (2003). Distinct HP1 and Su(var)3-9 complexes bind to sets of developmentally coexpressed genes depending on chromosomal location. *Genes and Development* **17**, 2825–2838.

Grewal, S. I. & Elgin, S. C. (2002). Heterochromatin: new possibilities for the inheritance of structure. *Current Opinion in Genetics and Development* **12**, 178–187.

Grimes, B., Rhoades, A. & Willard, H. (2002). α-Satellite DNA and vector composition influence rates of human artificial chromosome formation. *Molecular Therapy* **5**, 798–805.

Grover, D., Majumder, P. P., Rao, C. B., Brahmachari, S. K. & Mukerji, M. (2003). Non-random distribution of alu elements in genes of various functional categories: insight from analysis of human chromosomes 21 and 22. *Molecular Biology and Evolution* **20**, 1420–1424.

Hall, I. M., Shankaranarayana, G. D., Noma, K., Ayoub, N., Cohen, A. & Grewal, S. I. (2002). Establishment and maintenance of a heterochromatin domain. *Science* **297**, 2232–2237.

Hamilton, A. J. & Baulcombe, D. C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* **286**, 950–952.

Han, J. S., Szak, S. T. & Boeke, J. D. (2004). Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* **429**, 268–274

He, D. & Brinkley, B. R. (1996). Structure and dynamic organisation of centromeres/prekinetochores in the nucleus of mammalian cells. *Journal of Cell Science* **109**, 2693–2704.

Henderson, I. R., Owen, P. & Nataro, J. P. (1999). Molecular switches – the ON and OFF of bacterial phase variation. *Molecular Microbiology* **33**, 919–932.

Henikoff, S. (1996). Dosage-dependent modification of position-effect variegation in Drosophila. *Bioessays* **18**, 401–409.

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

Hodgson, J. W., Argiropoulos, B. & Brock, H. W. (2001). Site-specific recognition of a 70 base-pair element containing d(GA)n repeats mediates bithoraxoid polycomb group response element-dependent silencing. *Molecular and Cellular Biology* **21**, 4528–4543.

Hofnung, M. & Shapiro, J. (1999). *Research in Microbiology* **150** (special November–December double issue on bacterial repeats).

Hoskins, R. A., Smith, C. D., Carlson, J. W., Carvalho, A. B., Halpern, A., Kaminker, J. S., Kennedy, C., Mungall, C. J., Sullivan, B. A., Sutton, G. G., Yasuhara, J. C., Wakimoto, B. T., Myers, E. W., Celniker, S. E., Rubin, G. M. & Karpen, G. H. (2002). Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly. *Genome Biology* **3**, 0085.1–0085.16.

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

International Human Genome Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Ioshikhes, I. P. & Zhang, M. Q. (2000). Large-scale human promoter mapping using CpG islands. *Nature Genetics* **26**, 61–63.

Isobe, T., Yashino, M., Mizuno, K.-I., Lindahl, K. F., Koide, T., Gaudieri, S., Gojobori, T. & Shiroishi, T. (2002). Molecular characterization of the Pb recombination hotspot in the mouse major histocompatibility complex class II region. *Genomics* **80**, 229–235.

Jacob, F., Brenner, S. & Cuzin, F. (1963). On the regulation of DNA replication in Bacteria. *Cold Spring Harbor Symposium Quantitative Biology* **28**, 329–438.

Jacob, F. & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* **3**, 318–356.

Jeffreys, A. J., Monckton, D. G., Tamaki, K., Neil, D. L., Armour, J. A., MacLeod, A., Collick, A., Allen, M. & Jobling, M. (1993). Minisatellite variant repeat mapping: application to DNA typing and mutation analysis. *Experientia Supplementa* **67**, 125–139.

Jenuwein, T. (2002). An RNA-guided pathway for the epigenome. *Science* **297**, 2215–2218.

Jenuwein, T. & Allis, C. D. (2001). Translating the histone code. *Science* **293**, 1074–1080.

Jonsson, A. B., Nyberg, G. & Normark, S. (1991). Phase variation of gonococcal pili by frameshift mutation in pilC, a novel gene for pilus assembly. *EMBO Journal* **10**, 477–488.

Jordan, I. K., Rogozin, I. B., Glazko, G. V. & Koonin, E. V. (2003). Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends in Genetics* **19**, 68–72.

Kamath, R. S., Fraser, A. G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., Welchman, D. P., Zipperlen, P. & Ahringer, J. (2003). Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237.

Kapitonov, V. V. & Jurka, J. (2001). Rolling-circle transposons in eukaryotes. *Proceedings of the National Academy of Sciences, USA* **98**, 8714–8719.

Kashi, Y., King, D. & Doller, M. S. (1997). Simple sequence repeats as a source of quantitative variation. *Trends in Genetics* **13**, 74–78.

Khodarev, N. N., Bennett, T., Shearing, N., Sokolova, I., Koudelik, J., Walter, S., Villalobos, M. & Vaughn, A. T. M. (2000). LINE L1 retrotransposable element is targeted during the initial stages of apoptotic DNA fragmentation. *Journal of Cellular Biochemistry* **79**, 486–495.

Kinoshita, K. & Honjo, T. (2001). Linking class-switch recombination with somatic hypermutation. *Nature Reviews Molecular Cell Biology* **2**, 493–503.

Kirkness, E. F., Bafna, V., Halpern, A. L., Levy, S., Remington, K., Rusch, D. B., Delcher, A. L., Pop, M., Wang, W.,

Fraser, C. M. & Venter, J. C. (2003). The dog genome: survey sequencing and comparative analysis. *Science* **301**, 1898–1903.

Kitao, H., Arakawa, H., Kuma, K.-I., Yamagishi, H., Nakamura, N., Furusawa, S., Matsuda, H., Yasuda, M., Ekino, S. & Shimizu, A. (2000). Class switch recombination of the chicken IgH chain genes: implications for the primordial switch region repeats. *International Immunology* **12**, 959–968.

Kitano, H. (2002). Systems biology: a brief overview. *Science* **295**, 1662–1664.

Kloc, M. & Etkin, L. D. (1994). Delocalization of Vg1 mRNA from the vegetal cortex in *Xenopus* oocytes after destruction of Xlsirt RNA. *Science* **265**, 1101–1103.

Kropotov, A., Sedova, V., Ivanov, V., Sazeeva, N., Tomilin, A., Krutilina, R., Oei, S. L., Griesenbeck, J., Buchlow, G. & Tomilin, N. (1999). A novel human DNA-binding protein with sequence similarity to a subfamily of redox proteins which is able to repress RNA-polymerase-III-driven transcription of the Alu-family retroposons in vitro. *European Journal of Biochemistry* **260**, 336–346.

Kropotov, A. V., Yau, P., Bradbury, P. & Tomilin, N. (1997). Nonhistone chromatin proteins HMG1 and HMG2 stabilise one of the sequence-specific complexes, formed on the promoter of human retroposons of the ALU-family, by other nuclear proteins. *Molekuliarnaia Genetika, Mikrobiologia, I Virusologia* **4**, 32–36.

Kurenova, E., Champion, L., Biessmann, H. & Mason, J. M. (1998). Directional gene silencing induced by a complex subtelomeric satellite in Drosophila. *Chromosoma* **107**, 311–320.

Labrador, M. & Corces, V. G. (2002). Setting the boundaries of chromatin domains and nuclear organization. *Cell* **111**, 151–154.

Laemmli, U. K., Käs, E., Poljak, L. & Adachi, Y. (1992). Scaffold-associated regions: cis-acting determinants of chromatin structural loops and functional domains. *Current Opinion in Genetics and Development* **2**, 275–285.

Lebrun, E., Revardel, E., Boscheron, C., Li, R., Gilson, E. & Fourel, G. (2001). Protosilencers in *Saccharomyces cerevisiae* subtelomeric regions. *Genetics* **158**, 167–176.

Lewis, M., Chang, G., Horton, N. C., Kercher, M. A., Pace, H. C., Schumacher, M. A., Brennan, R. G. & Lu, P. (1996). Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science* **271**, 1247–1254.

Lim, J. K. & Simmons, M. J. (1994). Gross chromosome rearrangements mediated by transposable elements in Drosophila melanogaster. *Bioessays* **16**, 269–275.

Low, D. A., Weyand, N. J. & Mahan, M. J. (2001). Roles of DNA adenine methylation in regulating bacterial gene expression and virulence. *Infection and Immunity* **69**, 7197–7204.

Lynch, M. & Conery, J. (2003). The origins of genome complexity. *Science* **302**, 1401–1404.

Lyon, M. F. (2000). LINE-1 elements and X chromosome inactivation: a function for 'junk' DNA? *Proceedings of the National Academy of Sciences, USA* **97**, 6248–6249.

Lysnyansky, Y. R., Ron, Y. & Yogev, D. (2001). Juxtaposition of an active promoter to vsp genes via site-specific DNA inversions generates antigenic variation in *Mycoplasma bovis*. *Journal of Bacteriology* **183**, 5698–5708.

Macario, A. J. L., Lange, M., Ahring, B. K. & De Macario, E. C. (1999). Stress genes and proteins in the Archaea. *Microbiology and Molecular Biology Reviews* **63**, 923–967.

Mahmoudi, T., Katsani, K. R. & Verrijzer, C. P. (2002). GAGA can mediate enhancer function in *trans* by linking two separate DNA molecules. *EMBO Journal* **21**, 1775–1781.

Marczynski, G. T. & Shapiro, L. (1993). Bacterial chromosome origins of replication. *Current Opinion in Genetics and Development* **3**, 775–782.

Martienssen, R. A. (2003). Maintenance of heterochromatin by RNA interference of tandem repeats. *Nature Genetics* **35**, 213–214.

Matzke, M., Matzke, A. J. & Kooter, J. M. (2001). RNA: guiding gene silencing. *Science* **293**, 1080–1083.

Mazel, D., Dychinco, B., Webb, V. A. & Davies, J. (1998). A distinctive class of integron in the *Vibrio cholerae* genome. *Science* **280**, 605–608.

McKee, B. D., Hong, C. S. & Das, S. (2000). On the roles of heterochromatin and euchromatin in meiosis in *Drosophila*: mapping chromosomal pairing sites and testing candidate mutations for effects on X-Y nondisjunction and meiotic drive in male meiosis. *Genetica* **109**, 77–93.

Mette, M. F., Aufsatz, W., Van Der Winden, J., Matzke, M. A. & Matzke, A. J. (2000). Transcriptional silencing and promoter methylation triggered by double-stranded RNA. *EMBO Journal* **19**, 5194–5201.

Meyers, B. C., Tingey, S. V. & Morgante, M. (2001). Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Research* **11**, 1660–1676.

Miller, J. T., Dong, F., Jackson, S. A., Song, J. & Jiang, J. (1998). Retrotransposon-related DNA sequences in the centromeres of grass chromosomes. *Genetics* **150**, 1615–1623.

Miller, W. J., Nagel, A., Bachmann, J. & Bachmann, L. (2000). Evolutionary dynamics of the SGM transposon family in the *Drosophila obscura* species group. *Molecular Biology and Evolution* **17**, 1597–1609.

Miwa, Y., Nakata, A., Ogiwara, A., Yamamoto, M. & Fujita, Y. (2000). Evaluation and characterization of catabolite-responsive elements (cre) of *Bacillus subtilis*. *Nucleic Acids Research* **28**, 1206–1210.

Modrich, P. (1989). Methyl-directed DNA mismatch correction. *Journal of Biological Chemistry* **264**, 6597–6600.

Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H. Jr. (1999). Exon shuffling by L1 retrotransposition. *Science* **283**, 1530–1534.

Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562.

Mozer, B. A. & Benzer, S. (1994). Ingrowth by photoreceptor axons induces transcription of a retrotransposon in the developing *Drosophila* brain. *Development* **120**, 1049–1058.

Müller, A., Marins, M., Kamisugi, Y. & Meyer, P. (2002). Analysis of hypermethylation in the RPS element suggests a signal function for short inverted repeats in *de novo* methylation. *Plant Molecular Biology* **48**, 383–399.

Müller, J., Oehler, S. & Müller-Hill, B. (1996). Repression of lac promoter as a function of distance, phase and quality of an auxiliary lac operator. *Journal of Molecular Biology* **257**, 21–29.

Nagaki, K., Song, J., Stupar, R. M., Parokonny, A. S., Yuan, Q., Ouyang, S., Liu, J., Hsiao, J., Jones, K. M., Dawe, R. K., Buell, C. R. & Jiang, J. (2003*a*). Molecular and cytological analyses of large tracks of centromeric DNA reveal the structure and evolutionary dynamics of maize centromeres. *Genetics* **163**, 759–770.

Nagaki, K., Talbert, P. B., Zhong, C. X., Dawe, R. K., Henikoff, S. & Jiang, J. (2003*b*). Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of *Arabidopsis thaliana* centromeres. *Genetics* **163**, 1221–1225.

Nekrutenko, A. & Li, W.-H. (2001). Transposable elements are found in a large number of human protein coding regions. *Trends in Genetics* **17**, 619–625.

Nigumann, P., Redik, K., Matlik, K. & Speek, M. (2002). Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics* **79**, 628–634.

Norris, J., Fan, D., Aleman, C., Marks, J. R., Futreal, P. A., Wiseman, R. W., Inglehart, J. D., Deininger, P. L. & McDonnell, D. P. (1995). Identification of a new subclass of Alu DNA repeats which can function as estrogen receptor-dependent transcriptional enhancers. *Journal of Biological Chemistry* **270**, 22777–22782.

Novac, O., Matheos, D., Araujo, F. D., Price, G. B. & Zannis-Hadjopoulos, M. (2001). *In vivo* association of Ku with mammalian origins of DNA replication. *Molecular Biology of the Cell* **12**, 3386–3401.

Nurse, P., Masui, Y. & Hartwell, L. (1998). Understanding the cell cycle. *Nature Medicine* **4**, 1103–1106.

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

Ott, R. W. & Hansen, L. K. (1996). Repeated sequences from the *Arabidopsis thaliana* genome function as enhancers in transgenic tobacco. *Molecular and General Genetics* **252**, 563–571.

Pagel, M. & Johnstone, R. A. (1992). Variation across species in the size of the nuclear genome supports the junk-DNA explanation for the C-value paradox. *Proceedings of the Royal Society of London B: Biological Science* **249**(1325), 119–124.

Pal-Bhadra, M., Leibovitch, B. A., Gandhi, S. G., Rao, M., Bhadra, U., Birchler, J. A. & Elgin, S. C. R. (2004). Heterochromatic silencing and HP1 localization in *Drosophila* are dependent on the RNAi machinery. *Science* **303**, 669–672.

Pardue, M. L. & DeBaryshe, P. G. (2003). Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. *Annual Review of Genetics* **37**, 485–511.

Parish, D. A., Vise, P., Wichman, H. A., Bull, J. J. & Baker, R. J. (2002). Distribution of LINEs and other repetitive elements in the karyotype of the bat *Carollia*: implications for X-chromosome inactivation. *Cytogenetics and Genome Research* **96**, 191–197.

Parkhill, J., Achtman, M., James, K. D., Bentley, S. D., Churcher, C., Klee, S. R., Morelli, G., Basham, D., Brown, D., Chillingworth, T., *et al.* (2000). Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* **404**, 502–507.

Pelissier, T., Tutois, S., Tourmente, S., Deragon, J. M. & Picard, G. (1996). DNA regions flanking the major *Arabidopsis thaliana* satellite are principally enriched in Athila retroelement sequences. *Genetica* **97**, 141–151.

Pelisson, A., Mejlumian, L., Robert, V., Terzian, C. & Bucheton, A. (2002). *Drosophila* germline invasion by the endogenous retrovirus gypsy: involvement of the viral env gene. *Insect Biochemistry and Molecular Biology* **32**, 1249–1256.

Petrov, D. A. (2001). Evolution of genome size: new approaches to an old problem. *Trends in Genetics* **17**, 23–28.

Pimpinelli, S., Berloco, M., Fanti, L., Dimitri, P., Bonaccorsi, S., Marchetti, E., Caizzi, R., Caggese, C. & Gatti, M. (1995). Transposable elements are stable structural components of *Drosophila melanogaster* heterochromatin. *Proceedings of the National Academy of Sciences, USA* **92**, 3804–3808.

Ponger, L. & Mouchiroud, D. (2002). CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics* **18**, 631–633.

Prescott, D. M. (2000). Genome gymnastics: unique modes of DNA evolution and processing in ciliates. *Nature Reviews Genetics* **1**, 191–198.

Price, G. B., Allarakhia, M., Cossons, N., Nielsen, T., Diaz-Perez, M., Friedlander, P., Tao, L. & Zannis-Hadjopoulos, M. (2003). Identification of a *cis*-element that determines autonomous DNA replication in eukaryotic cells. *Journal of Biological Chemistry* **278**, 19649–19659.

Pryde, F. E. & Louis, E. J. (1999). Limitations on natural TPE in yeast. *EMBO Journal* **18**, 2538–2550.

Ptashne, M. (1986). *A Genetic Switch: Phage lambda and Higher Organisms*. Cell Press and Blackwell Scientific Publications, Cambridge, MA, 2nd edition.

Quiberoni, A., Rezaiki, L., El Karoui, M., Biswas, I., Tailliez, P. & Gruss, A. (2001). Distinctive features of homologous recombination in an 'old' microorganism, *Lactococcus lactis*. *Research in Microbiology* **152**, 131–139.

Reinhart, B. J. & Bartel, D. P. (2002). Small RNAs correspond to centromere heterochromatic repeats. *Science* **297**, 1831.

Richardson, A. R. & Stojiljkovic, I. (1999). HmbR a hemoglobin-binding outer membrane protein of *Neisseria meningitidis* undergoes phase variation. *Journal of Bacteriology* **18**, 2067–2074.

Robertson, K. D. (2002). DNA methylation and chromatin: unraveling the tangled web. *Oncogene* **21**, 5361–5379.

Rollini, P., Namciu, S. J., Marsden, M. D. & Fournier, R. E. K. (1999). Identification and characterization of nuclear matrix-attachment regions in the human serpin gene cluster at 14q32.1. *Nucleic Acids Research* **27**, 3779–3791.

Rothenburg, S., Koch-Nolte, F., Rich, A. & Haag, F. (2001). A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. *Proceedings of the National Academy of Sciences, USA* **98**, 8985–8990.

Roy, P. J., Stuart, J. M., Lund, J. & Kim, S. K. (2002). Chromosomal clustering of muscle expressed genes in *Caenorhabditis elegans*. *Nature* **418**, 975–979.

Rubin, C. M., Kimura, R. H. & Schmid, C. W. (2002). Selective stimulation of translational expression by Alu RNA. *Nucleic Acids Research* **30**, 3253–3261.

Rui, W. J. (1999). Regulation of eukaryotic DNA replication and nuclear structure. *Cell Research* **9**, 163–170.

Sangwan, I. & O'Brian, M. R. (2002). Identification of a soybean protein that interacts with GAGA element dinucleotide repeat DNA. *Plant Physiology* **129**, 1788–1794.

Santi, L., Wang, Y., Stile, M. R., Berendzen, K., Wanke, D., Roig, C., Pozzi, C., Muller, K., Muller, J., Rohde, W. & Salamini, F. (2003). The GA octodinucleotide repeat binding factor BBR participates in the transcriptional regulation of the homeobox gene Bkn3. *Plant Journal* **34**, 813–826.

Sarkari, J., Pandit, N., Moxon, E. R. & Achtman, M. (1994). Variable expression of the Opc outer membrane protein in *Neisseria meningitidis* is caused by size variation of a promoter containing poly-cytidine. *Molecular Microbiology* **13**, 207–217.

Saunders, N. J., Jeffries, A. C., Peden, J. F., Hood, D. W., Tettelin, H., Rappuoli, R. & Moxon, E. R. (2000). Repeat-associated phase variable genes in the complete genome sequence of *Neisseria meningitidis* strain MC58. *Molecular Microbiology* **37**, 207–215.

Saveliev, A., Everett, C., Sharpe, T., Webster, Z. & Festenstein, R. (2003). DNA triplet repeats mediate

heterochromatin-protein-1-sensitive variegated gene silencing. *Nature* **422**, 909–913.

Sawitzke, J. & Austin, S. (2001). An analysis of the factory model for chromosome replication and segregation in bacteria. *Molecular Microbiology* **40**, 786–794.

Schild-Poulter, C., Matheos, D., Novac, O., Cui, B., Giffin, W., Ruiz, M. T., Price, G. B., Zannis-Hadjopoulos, M. & Hache, R. J. (2003). Differential DNA binding of Ku antigen determines its involvement in DNA replication. *DNA and Cell Biology* **22**, 65–78.

Schmid, C. (1996). Alu: structure, origin, evolution, significance, and function of one-tenth of human DNA. *Progress in Nucleic Acids Research and Molecular Biology* **53**, 283–319.

Scholes, D. T., Kenny, A. E., Gamache, E. R., Mou, Z. & Curcio, M. J. (2003). Activation of a LTR-retrotransposon by telomere erosion. *Proceedings of the National Academy of Sciences, USA* **100**, 15736–15741.

Schotta, G., Ebert, A., Dorn, R. & Reuter, G. (2003). Position-effect variegation and the genetic dissection of chromatin regulation in Drosophila. *Seminars in Cell and Developmental Biology* **14**, 67–75.

Schueler, M. G., Higgins, A. W., Rudd, M. K., Gustashaw, K. & Willard, H. F. (2001). Genomic and genetic definition of a functional human centromere. *Science* **294**, 109–115.

Selker, E. U. (1990). Premeiotic instability of repeated sequences in *Neurospora crassa*. *Annual Review of Genetics* **24**, 579–613.

Shapiro, J. A. (1983). *Mobile Genetic Elements*. Academic Press, New York.

Shapiro, J. A. (1999*a*). Genome system architecture and natural genetic engineering in evolution. *Annals of the New York Academy of Science* **870**, 23–35.

Shapiro, J. A. (1999*b*). Transposable elements as the key to a 21st Century view of evolution. *Genetica* **107**, 171–179.

Shapiro, J. A. (2002*a*). Genome organisation and reorganisation in evolution: formatting for computation and function. In *From Epigenesis to Epigenetics: The Genome in Context* (eds. L. Van Speybroeck, G. Van de Vijver and D. De Waele). *Annals of the New York Academy of Science* **981**, 111–134.

Shapiro, J. A. (2002*b*). Repetitive DNA, genome system architecture and genome reorganisation. *Research in Microbiology* **150**, 447–453.

Sharp, S. I., Pickrell, J. K. & Jahn, C. L. (2003). Identification of a novel 'chromosome scaffold' protein that associates with *Tec* elements undergoing *en masse* elimination in *Euplotes crassus*. *Molecular Biology of the Cell* **14**, 571–584.

Shiroishi, T., Koide, T., Yoshino, M., Sagai, T. & Moriwaki, K. (1995). Hotspots of homologous recombination in mouse meiosis. *Advances in Biophysics* **31**, 119–132.

Silva, J. C., Shabalina, S. A., Harris, D. G., Spouge, J. L. & Kondrashovi, A. S. (2003). Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR-and L2-derived sequences within the mouse and human genomes. *Genetical Research* **82**, 1–18.

Skryabin, B. V., Kremerskothen, J., Vassilacopoulou, D., Disotell, T. R., Kapatinov, V. V., Jorka, J. & Brosius, J. (1998). The BC200 RNA gene and its neural expression are conserved in Anthropoidea (primates). *Journal of Molecular Evolution* **47**, 677–685.

Song, X., Sui, A. & Garen, A. (2004). Binding of mouse VL30 retrotransposon RNA to PSF protein induces genes repressed by PSF: Effects on steroidogenesis and oncogenesis. *Proceedings of the National Academy of Sciences, USA* **101**, 621–626.

Speek, M. (2001). Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Molecular and Cellular Biology* **21**, 1973–1985.

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

Spofford, J. B. & DeSalle, R. (1991). Nucleolus organiser-suppressed position-effect variegation in *Drosophila melanogaster*. *Genetical Research* **57**, 245–255.

Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., *et al.* (2003). The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biology* **1**, 166–192.

Stern, A. & Meyer, T. F. (1987). Common mechanism controlling phase and antigenic variation in pathogenic neisseriae. *Molecular Microbiology* **1**, 5–12.

Tchénio, T., Casella, J.-F. & Heidmann, T. (2000). Members of the SRY family regulate the human LINE retrotransposons. *Nucleic Acids Research* **28**, 411–415.

Tomilin, N. V., Bozhkov, V. M., Bradbury, E. M. & Schmid, C. W. (1992). Differential binding of human nuclear proteins to Alu subfamilies. *Nucleic Acids Research* **20**, 2941–2945.

Tovar, D., Faye, J. C. & Favre, G. (2003). Cloning of the human RHOB gene promoter: characterization of a VNTR sequence that affects transcriptional activity. *Genomics* **81**, 525–530.

Trifonov, E. N. (1999). Elucidating sequence codes: three codes for evolution. *Annals of the New York Academy of Science* **870**, 330–338.

Tsuchiya, T., Saëgusa, Y., Taira, T., Mimori, T., Iguchi-Ariga, S. M. M. & Ariga, H. (1998). Ku antigen binds to Alu family DNA. *Journal of Biochemistry* **123**, 120–127.

Vafa, O. & Sullivan, K. F. (1997). Chromatin containing CENP-A and alpha-satellite DNA is a major component of the inner kinetochore plate. *Current Biology* **7**, 897–900.

van der Ende, A., Hopman, C. T., Zaat, S., Essink, B. B., Berkhout, B. & Dankert, J. (1995). Variable expression of class 1 outer membrane protein in *Neisseria meningitidis* is caused by variation in the spacing between the −10 and −35 regions of the promoter. *Journal of Bacteriology* **177**, 2475–2480.

van Driel, R., Fransz, P. F. & Verschure, P. J. (2003). The eukaryotic genome: a system regulated at different hierarchical levels. *Journal of Cell Science* **116**, 4067–4075.

Vansant, G. & Reynolds, W. F. (1995). The consensus sequence of a major Alu subfamily contains a functional retinoic acid response element. *Proceedings of the National Academy of Sciences, USA* **92**, 8229–8233.

Van Speybroeck, L., Van de Vijver, G. & De Waele, D. (2002). *From Epigenesis to Epigenetics: The Genome in Context*. Annals of the New York Academy of Science, volume 981, New York.

van Steensel, B., Delrow, J. & Bussemaker, H. J. (2003). Genomewide analysis of *Drosophila* GAGA factor target genes reveals context-dependent DNA binding. *Proceedings of the National Academy of Sciences, USA* **100**, 2580–2585.

Vassetzky, N. S., Ten, O. A. & Kramerov, D. A. (2003). B1 and related SINEs in mammalian genomes. *Gene* **319**, 149–160.

Verdel, A., Jia, S., Gerber, S., Sugiyama, T., Gygi, S., Grewal, S. I. S. & Moazed, D. (2004). RNAi-mediated targeting of heterochromatin by the RITS complex. *Science* **303**, 672–676.

Versteeg, R., van Schaik, B. D., van Batenburg, M. F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H. J. & van Kampen, A. H. (2003). The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat

pattern for domains of highly and weakly expressed genes. *Genome Research* **13**, 1998–2004.

Vinogradov, A. E. (2003). Selfish DNA is maladaptive: evidence from the plant Red List. *Trends in Genetics* **19**, 609–614.

Volpe, T. A., Kidner, C., Hall, I. M., Teng, G., Grewal, S. I. & Martienssen, R. A. (2002). Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297**, 1833–1837.

Volpe, T. A., Schramke, V., Hamilton, G. L., White, S. A., Teng, G., Martienssen, R. A. & Allshire, R. C. (2003). RNA interference is required for normal centromere function in fission yeast. *Chromosome Research* **11**, 137–146.

Wahle, E. & Ruegsegger, U. (1999). 3′-End processing of pre-mRNA in eukaryotes. *FEMS Microbiology Reviews* **23**, 277–295.

Wakimoto, B. T. (1998). Beyond the nucleosome: epigenetic aspects of position-effect variegation in Drosophila. *Cell* **93**, 321–324.

Wang, G., van Dam, A. P. & Dankert, J. (1995). Analysis of a VMP-like sequence (vls) locus in *Borrelia garinii* and Vls homologues among four *Borrelia burgdorferi sensu lato* species. *FEMS Microbiological Letters* **199**, 39–45.

Watson, J. B. & Sutcliffe, J. G. (1987). Primate brain-specific cytoplasmic transcript of the Alu repeat family. *Molecular and Cellular Biology* **7**, 3324–3327.

Weiner, A. M., Deininger, P. L. & Efstratiadis, A. (1986). Nonviral retroposons: Genes, pseudogenes and transposable elements generated by the reverse flow of genetic information. *Annual Review of Biochemistry* **55**, 631–661.

Widlund, H. R., Kuduvalli, P. N., Bengtsson, M., Cao, H., Tullius, T. D. & Kubista, M. (1999). Nucleosome structural features and intrinsic properties of the TATAAACGCC repeat sequence. *Journal of Biological Chemistry* **274**, 31847–31852.

Wyngaard, G. A. & Gregory, T. R. (2001). Temporal control of DNA replication and the adaptive value of chromatin diminution in copepods. *Journal of Experimental Zoology* **291**, 310–316.

Wyrick, J. J., Aparicio, J. G., Chen, T., Barnett, J. D., Jennings, E. G., Young, R. A., Bell, S. P. & Aparicio, O. M. (2001). Genome-wide distribution of ORC and MCM proteins in *S. cerevisiae*: high-resolution mapping of replication origins. *Science* **14**, 2357–2360.

Yang, Z., Bofelli, D., Boonmark, N., Schwartz, K. & Lawn, R. (1998). Apolipoprotein(a) gene enhancer resides within a LINE element. *Journal of Biological Chemistry* **273**, 891–897.

Yates, P. A., Burman, R. W., Mummaneni, P., Krussel, S. & Turker, M. S. (1999). Tandem B1 elements located in a mouse methylation center provide a target for *de novo* DNA methylation. *Journal of Biological Chemistry* **274**, 36357–36361.

Youn, B. S., Lim, C. L., Shin, M. K., Hill, J. M. & Kwon, B. S. (2002). An intronic silencer of the mouse perforin gene. *Molecular Cells* **13**, 61–68.

Yu, J., Hu, S., Wang, J., Wong, G. K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., *et al.* (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79–92.

Yuh, C. H., Bolouri, H. & Davidson, E. H. (1998). Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* **279**, 1896–1902.

Zaiss, D. M. W. & Kloetzel, P.-M. (1999). A second gene encoding the mouse proteosome activator b subunit is part of a LINE1 element and is driven by a LINE1 promoter. *Journal of Molecular Biology* **287**, 829–835.

Zearfoss, N. R., Chan, A. P., Kloc, M., Allen, L. H. & Etkin, L. D. (2003). Identification of new Xlsirt family members in the *Xenopus laevis* oocyte. *Mechanisms of Development* **120**, 503–509.

Zhong, C. X., Marshall, J. B., Topp, C., Mroczek, R., Kato, A., Nagaki, K., Birchler, J. A., Jiang, J. & Dawe, R. K. (2002). Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. *Plant Cell* **14**, 2825–2836.

Zhou, Y.-H., Zheng, J. B., Gu, X., Li, W.-H. & Saunders, G. F. (2000). A novel Pax-6 binding site in rodent B1 repetitive elements: coevolution between developmental regulation and repeated elements? *Gene* **245**, 319–328.

Zhou, Y.-H., Zheng, J. B., Gu, X., Saunders, G. F. & Yung, W.-K. A. (2002). Novel Pax6 binding sites in the human genome and the role of repetitive elements in the evolution of gene regulation. *Genome Research* **12**, 1716–1722.

Zhu, L., Swergold, G. D. & Seldin, M. F. (2003). Examination of sequence homology between human chromosome 20 and the mouse genome: intense conservation of many genomic elements. *Human Genetics* **113**, 60–70.

Zuckerkandl, E. (2002). Why so many noncoding nucleotides? The eukaryote genome as an epigenetic machine. *Genetica* **115**, 105–129.