Thinking about evolution in terms of cellular computing

JAMES A. SHAPIRO
Department of Biochemistry and Molecular Biology, University of Chicago,
5614 S. Dorchester Ave., Chicago, IL 60637, USA (E-mail: jsha@uchicago.edu)

Abstract. The past five decades of molecular genetics have produced many discoveries about genome structure and function that can only be understood from an informatic perspective:

– distinct sequence codes to mark the individual steps in packaging, expression, replication, transmission, repair and restructuring of DNA molecules;
– modularity of data files for RNA and protein products;
– combinatoric organization of signals to format the genome for differential functioning during cellular and organismal cycles;
– direct participation of DNA in the execution of biological algorithms (formation of highly structured nucleoprotein complexes);
– hierarchical organization of genomic subsystems to form higher level system architectures.

This review will discuss aspects of genome organization and genome change that require a more formal computational analysis. We will see how modern results indicate that genome evolution has many similarities to computer system engineering. The ability of cells to control the function of natural genetic engineering systems is central to the genome's potential as a Read–Write information storage system.

Key words: adaptive mutation, genome system architecture, mobile genetic elements, natural genetic engineering, retrotransposon, transposon

Abbreviations: CSR – class switch recombination; DGR – diversity generating retroelement; DSB – double-strand break; DTE – DNA transposable element; HERV – human endogenous retrovirus; LINE – long interspersed nucleotide element; MGE – mobile genetic element; MLV – murine leukemia virus; NHEJ – non-homologous end joining; V(D)J, the repeated protein cassettes of the vertebrate immune system; RSS – recombination signal sequence; RTE – RNA transposable element or retrotransposon; SINE – short interspersed nucleotide element

1. Introduction: thinking about DNA as a data storage medium

It is widely recognized that DNA sequences contain information about protein and RNA structure. By interacting with other molecules

in the cell, DNA stores information for various periods. We can distinguish three time scales for data storage by DNA molecules:

- many organismal generations (genetic storage in primary sequence);
- multiple cell generations (epigenetic storage or "imprinting" *via* DNA modifications and chromatin configurations);
- within a single cell cycle (short-term regulatory storage in metastable nucleoprotein complexes).

Genome sequencing examines the long-term primary structure of DNA. Analysis of epigenetic inheritance deals with intermediate-term storage that is critical for maintaining patterns of cellular differentiation and multicellular development (Jenuwein, 2002; Van Driel et al., 2003). Epigenetic inheritance relies upon chemical modification of DNA in ways that do not alter sequence content but which affect access to specific regions of the genome. Short-term storage of data about physiology, DNA replication, or progress through the cell cycle involves DNA-protein-RNA complexes that are subject to rapid modification by transient changes in cellular signaling molecules.

The core argument of this paper is that information storage at all three time scales can best be viewed from a computational perspective. By "computational," I refer to processes that involve the evaluation of multiple inputs and the subsequent working out of decisions leading to outputs that may be predictable but are not automatic. Cellular computing operates through networks of stereospecific molecular interactions and employs combinatorial principles to arrive at its decisions (*cf.* Bray, 1990, 1995; Hartwell et al., 1999). We are rapidly accumulating knowledge about individual interactions and the detailed structure of some networks (Milo et al., 2004), but it is clear that we do not yet have a developed conceptual understanding of overall network architectonics or the basic principles of cellular computation.

In terms of understanding genome organization as a cellular data storage system, we need to distinguish between the content of data files (RNA and protein coding sequences) and the signals needed to format the genome for involvement in many cellular functions: data file access, genome replication and transmission to progeny cells, packaging and physical organization of DNA, repair and error correction, and genome restructuring. Although a large majority of biologists would currently argue that genome restructuring is not a cellular function, the informatics metaphor tells us that a read-write

storage system is vastly more powerful than a read-only memory. There is overwhelming evidence that cells possess multiple natural genetic engineering systems to restructure their genomes (Bukhari et al., 1977; Shapiro, 1983, 1992; Craig et al., 2002). As we shall see below, these systems have the potential to write new information in a biologically regulated manner.

2. Generic formatting signals and the repetitive component of the genome

In order to function in concert with the cellular apparatuses for general genomic processes such as transcription, replication, packaging and repair, DNA formatting signals need to have a much smaller information content than the data files. The signals must be generic in nature and used repeatedly for the same function. This is the computational reason that genomes contain repetitive DNA (Sternberg and Shapiro, 2005; Shapiro and Sternberg, 2005). As is well known from studies of transcriptional regulation, tremendous computational specificity can be achieved through different combinations of generic signals (Arnone and Davidson, 1997; Yuh et al., 1998; Bolouri and Davidson, 2002).

A complementary fundamental reason for repetitive DNA signals resides in the way cells carry out most stereospecific molecular interactions involving the genome. Our understanding of this process is based upon the pioneering studies of bacterial regulation and DNA replication. The definition of *cis*-acting sites like operators, promoters and replication origins added a new class of determinant to genetics, distinct from the "structural genes" that encode a specific product (Jacob and Monod, 1961; Jacob et al., 1963). Over the years, we have learned how these sites are iterated so that weak individual protein-DNA binding interactions can synergize with protein-protein binding (Ptashne, 1986). Such repetition and cooperativity provide the thermodynamic stability for nucleo-protein machines to operate effectively in transcription and replication. The algorithmic (computational) details of cooperativity between related and different *cis*-acting signals in the paradigmatic *lac* operon system have already been described (Shapiro, 2002b; Shapiro and Sternberg, 2005). The same principles of cooperativity apply in other classic bacterial examples, such as phage lambda

regulation and control of the SOS DNA repair system, and they can be generalized to more complex eukaryotic cells (Ptashne, 1986).

Small *cis*-acting signals (≤100 base pairs) constitute only one class of genomic repeats. Virtually all genomes contain repetitive DNA comprising from 5% to over 80% of the total genome (tabulated in Shapiro and Sternberg, 2005). Much of this DNA is larger repeat elements that range in size from hundreds to tens of thousands of base pairs. These larger repeats have defined structures. Detailed study of any particular repeat has inevitably revealed that it contains multiple generic signals, and many also contain coding sequences. In the genomes of animals and plants so far sequenced, the repetitive component exceeds the protein-coding component, often by more than an order of magnitude. The human genome, for example, contains less than 3% protein-coding exons but over 50% repetitive DNA. About 40% of the human genome consists of mobile genetic elements (MGEs) and their remnants dispersed throughout the chromosome arms (International Human Genome Consortium, 2001). Another 18% of human DNA consists of tandemly repeated elements that cluster around the centromeres of each
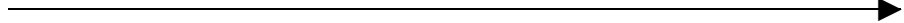
Figure 1. Some rearrangements of external DNA by mobile genetic elements. (a) DNA transposons that undergo replicative transposition in bacteria (Shapiro, 1979) can form deletions and excisions and can then fuse the excised circle with another region of the genome. The resulting segment of DNA flanked by two copies of the transposon is itself mobile and can insert, along with its flanking transposons, into yet another site in the genome, They can also generate reciprocal translocations, with each recombinant duplex carrying a copy of the replicated transposon. (b) DNA transposons that undergo non-replicative transposition in diploid organisms, such as P factors in Drosophila (http://engels.genetics.wisc.edu/Pelements/index.html), can sometimes cause rearrangements that involve two ends of different elements on sister chromatids. In such cases, the two ends can carry out strand transfer reactions at a site on a homologous chromosome to generate deletions and tandem duplications. (c) Long interspersed nucleotide elements (LINEs) can promote the target-primed reverse transcription of a spliced messenger RNA molecule to insert an intron-free copy of the coding sequence into a new genomic location. Sometimes transcription from a LINE element passes the normal polyA addition site and extends into 3′ flanking DNA. This extended transcript can be partially reverse-transcribed, and the 5′ truncated cDNA copy can be inserted into a new genomic location. If the transduced adjacent sequence carries an exon, and if the insertion occurs in a genetic locus, then the transduced exon has been added to the coding sequences of the target locus, providing a mechanism for domain accretion, as illustrated by Figure 42 in International Human Genome Consortium (2001).

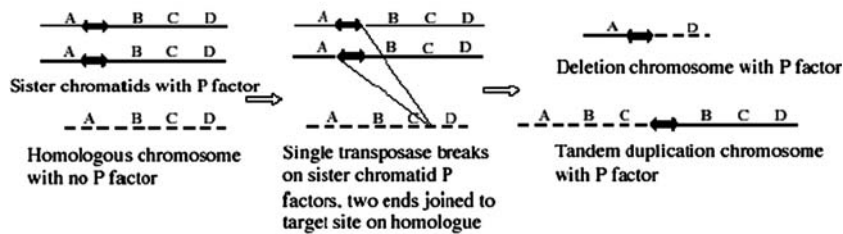chromosome and contain signals for binding centromere-specific proteins.

3. Mobile genetic elements and non-random genome restructuring

MGEs can be categorized in several ways (Shapiro, 1983; Craig et al., 2002; Shapiro and Sternberg, 2005). The diversity is important
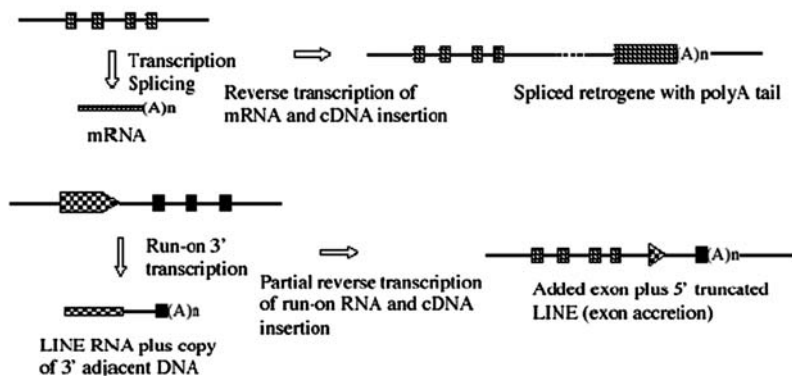
(a) Replicative Transposition:
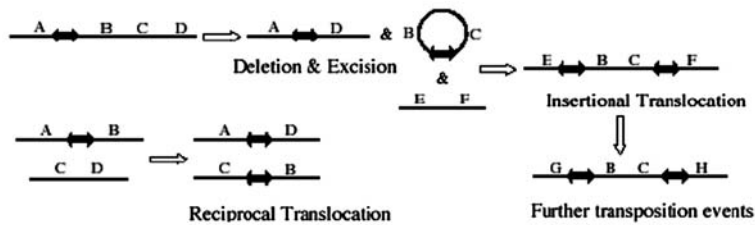
(b) Non-Replicative Transposition:

(c) Target-Primed Reverse Transcription:

because different MGE classes restructure the genome in different ways (*e.g.* Figure 1). The most important categories relate to whether MGEs are capable of encoding their own mobility from one genomic location to another (transposition) and to their mechanism of mobility (*cf.* Figure 1 of Kazazian, 2004). If they encode the necessary proteins, they are said to be "autonomous" elements. "Non-autonomous" MGEs depend upon related autonomous elements for movement to new genomic locations. The mobility mechanisms break down into DNA-based transposition (DNA transposable elements or DTEs) and transposition through an RNA intermediate *via* reverse transcription (retrotransposons or RTEs). Both the DTEs and RTEs can be further subdivided by their particular mechanisms of transposition and retrotransposition (reviewed in Craig et al., 2002). DTEs can transpose by a replicative mechanism, found in a subset of bacterial elements, or by non-replicative mechanisms involving double-strand breaks at the ends of the DTE. Retroviral-like RTEs depend upon flanking structures called long terminal repeats (LTRs) for their reverse transcription, while other RTEs have no terminal repeats (*cf.* Figure 2 of Kazazian, 2004). The non-LTR RTEs are divided into autonomous "long interspersed nucleotide elements" (LINEs) and non-autonomous "short interspersed nucleotide elements" (SINEs). LINE and SINE reverse transcription is a promiscuous process because it does not depend upon any feature of the RNA other than the poly-A tail found at the end of many transcripts (Kazazian, 2000, 2004; Figure 1).
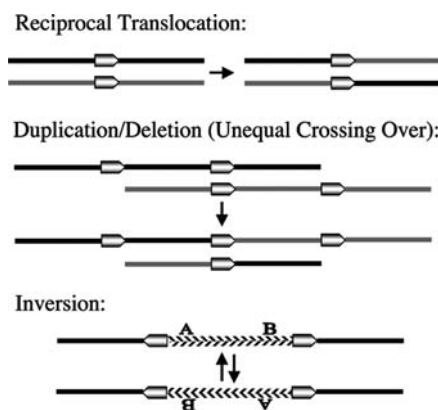

Figure 2. Consequences of ectopic homologous recombination between dispersed repeats.

For the purposes of thinking computationally about genome organization and reorganization, there are two general points to emphasize about MGEs:

1. Each element constitutes a defined ensemble of signals that can affect various aspects of genome function, such as transcription, chromatin formatting and/or DNA replication. Wherever a particular MGE is located, it consequently influences the structure and function of the surrounding DNA in a complex but reproducible fashion. Thus, MGEs can be considered mobile regulatory modules capable of acting at diverse locations throughout the genome (Shapiro, 2005b). There is considerable evidence that distribution of MGEs has served to establish novel regulatory configurations at individual loci (Britten, 1996; Brosius, 1999; Jordan et al., 2003) or at sets of coregulated loci (Peaston et al., 2004). In addition, MGEs also provide a source of new sequence components for protein evolution (Nekrutenko and Li, 2001).

2. Genetic mobility does not only apply to MGEs but also to other regions of the genome. From the earliest days of studying mobile genetic elements, it was apparent that they could serve as general agents of genome restructuring (Shapiro, et al., 1977). Each class of MGE has its own characteristic modality of mobilizing external DNA (Figure 1). DTEs, for example, rearrange large genome segments, often in the megabase range, to generate deletions, duplications, inversions, and insertional translocations (Shapiro, 1979, 2005a; http://engels.genetics.wisc.edu/Pelements/index.html; Harden and Ashburner, 1990). LINE elements, on the other hand, mobilize short segments of adjacent 3' sequence, a process that has probably played a key role in domain shuffling during protein evolution (Moran et al., 1999; Kazazian, 2000). In addition, LINE element functions can reverse transcribe and integrate mRNA sequences into the genome to generate retrogenes, as in the amplification of olfactory receptors (Brosius, 1999; Figure 1). Naturally, distributed copies of all MGEs can also play the roles of dispersed sequence homologies and lead to chromosome rearrangements by ectopic homologous recombination (Kazazian, 2000, 2004; Bailey et al., 2003; Figure 2).

Informatically, these two aspects of MGE action can be expressed as (1) the ability to attach pre-programmed control routines to new execution functions and (2) the capacity to amplify

and/or join together any two portions of the genomic program. MGEs thus represent an important part of the cellular capacity for genome system engineering. Like its human analogues, natural genetic engineering involves trial and error but is definitely not a random process. Movement of well-defined MGE structures to new locations and rearrangement of external DNA in predictable ways by particular MGEs, as illustrated in Figure 1, are non-random activities. Moreover, as we shall see below, cellular computing networks regulate when MGEs become active and where they act in the genome.

4. Cellular capacities for natural genetic engineering at short time scales – the ciliate example

A key issue in evolutionary theory is the time needed to effect significant genome restructuring to encode the expression of novel adaptive functions. With the realization that cells contain MGEs and other biochemical agents of DNA restructuring (nucleases. ligases, polymerases, etc.), we can conceive rapid genomic changes when those functions act at high levels. Do they actually do so, or (as conventional theory assumes) is the process of DNA restructuring so destructive that it can only proceed in small steps? The answer is that many organisms display short-term bursts of MGE activity or rapid genome rearrangement. The most spectacular examples are found among the ciliated protozoa (Prescott, 2000).

Ciliated protozoa (ciliates) are single-celled eukaryotic organisms that have separate germ-line and somatic genomes in two morphologically and functionally distinct nuclei: the micronucleus and the macronucleus. The small micronucleus undergoes meiosis to form haploid gametes prior to mating, while the much larger macronucleus encodes cellular RNA and protein molecules. The micronuclear genome comprises a small number of long chromosomes containing many DNA sequences absent from the macronuclear genome, which consists of many copies of numerous small telomere-capped fragments, each mini-chromosome containing a single genetic locus. We know the macronucleus encodes all essential functions because some species of ciliates have no micronucleus and reproduce in a strictly vegetative manner. Although the ciliate macronuclear genome system architecture is radically different from that of most eukaryotes, it does share some key features, such as telomerase and self-splicing RNA

molecules, both of which were discovered in these fascinating organisms (Kruger et al., 1982; Greider and Blackburn, 1987).

When ciliates encounter starvation conditions, they undergo meiosis and conjugal mating to produce cells containing new zygote micronuclei. In a process that takes several hours but is complete before any cell division occurs, the zygote macronucleus divides, the old micro- and macro-nuclei degenerate, and one sibling zygote micronucleus develops into a new functional macronucleus. This process involves massive natural genetic engineering, including enodreplication of micronuclear chromosomes, excision and degradation of over 90% of the micronuclear genome, formation of DNA segments encoding individual RNA and protein molecules, and capping of these mini-chromosomes with telomeres.

Some of the massive DNA excision events in macronucleus development resemble double-strand breakage steps at specific signals also observed in the movement of DNA transposons (Jahn and Klobutcher, 2002). But macronuclear development is more than just a series of DNA cleavage events. It involves true genetic engineering; separate DNA fragments join to construct protein coding sequences. The micronuclear genome contains the coding information for certain proteins in discrete, scrambled segments (sometimes more than 40 for a single protein). These segments must be excised, aligned and joined precisely to form functional macronuclear mini-chromosomes. Accomplishing this task reliably for many different protein-coding segments after each mating cycle demands thousands of highly controlled DNA cleavage, recognition and subsequent ligation steps. The biocomputational implications of DNA unscrambling have been discussed by Landweber and her colleagues (Landweber and Kari, 1999; Ruben and Landweber, 2000).

We are still ignorant of many aspects of how ciliates control DNA rearrangements during macronuclear development. However, one experimental finding deserves special mention because it reveals an unexpected capacity for communication and memory in DNA restructuring. DNA segments present in the ''old'' macronucleus before mating can determine the structure of the corresponding segments that develop in the ''new'' macronucleus (summarized in Meyer and Garnier, 2002; Jahn and Klobutcher, 2002). Two types of effects have been documented. In one case, DNA deleted from the old macronucleus but present in the zygote micronucleus is also missing from the new macronucleus. In the other case, presence of a normally excised

segment in the old macronucleus leads to retention of that segment in the new macronucleus. Somehow, the degenerating old macronucleus communicates its DNA structure to the developing new macronucleus. The micronuclear (germ-line) genome does not change, but the macronuclear (somatic) genome retains modifications from one generation to the next. This is a kind of "epigenetic" inheritance, but it likely is quite different in mechanism from chromatin-based epigenetic inheritance in multicellular eukaryotes.

Ciliate macronuclear development provides a striking counter-example to the limits placed on processes of DNA change by conventional evolutionary theories. In particular, the ciliates show how cells can completely disassemble the micronuclear genome and efficiently reassemble thousands of fragments into a functional macronuclear genome in a single cell generation. This process also demonstrates capacities for tight regulation of natural genetic engineering functions (they are only active after mating) and for communicating genome structural information between different cell compartments.

5. Cellular capacities for natural genetic engineering at intermediate time scales – rapid protein evolution in the immune system

There are many cases where genome rearrangements occur in defined ways linked to cellular differentiation over the course of multicellular development (Muller and Tobler, 2000; Goday and Esteban, 2001; Redi et al., 2001). The most intensively studied are the highly orchestrated changes that assemble and then optimize the DNA sequences encoding antigen recognition molecules, particularly antibodies, in differentiating lymphocytes (Bassing et al., 2002; Gellert, 2002; Mosto-slavsky et al., 2003).

The lymphocytes have to solve an extremely difficult computational/evolutionary problem: how to evolve a virtually infinite array of recognition molecules based on finite coding sequence resources. This limitless array is needed to detect and destroy an unpredictable variety of foreign invaders. Lymphocytes accomplish this task with a sequence of non-random but flexible DNA rearrangements coupled to a positive feedback loop amplifying those individual cells that have succeeded in producing the right antigen-binding specificity (clonal selection; Burnet, 1964). Once amplified, these cells undergo further non-random DNA changes to improve their binding specificity (somatic hypermutation; Kinoshita and Honjo, 2001; Franklin and

Blanden, 2004) and to alter the antibody class so that the antigen recognition molecules are directed to the right place in the body (class switch recombination or CSR; Kinoshita and Honjo, 2001; Chaudhuri and Alt, 2004).

Lymphocyte DNA rearrangements are full of important lessons about cellular natural genetic engineering capabilities and how cells regulate them (Shapiro, 2005a). Readers interested in the molecular and cellular details should consult the reviews cited above. The most important lesson is that particular cells, lymphocytes, have evolved to evolve specific adaptations rapidly. Thus, there can be no fundamental barrier to the evolution of more efficient evolution mechanisms. Among the other lessons we can draw from antibody formation are the following:

1. Lymphocytes can direct DNA rearrangements to specific sites in the genome by several different mechanisms:

   – Combinatorial diversity in antibody synthesis results from a DNA transposon-related activity making double-strand breaks (DSBs) at specific recombination signal sequences (RSSs). The RSSs are located next to variable (V), diversity (D), and joining (J) coding segments so that V–J and V–D–J joining events can occur.
   – In CSR, specific switch (S) regions are activated for breakage and rejoining by transcription from lymphokine-regulated promoters. Different lymphokine signals thus lead to the formation of different classes of antibody with a particular binding specificity. It is important to emphasize that the connection between transcription and DNA rearrangement provides a mechanistic basis for computational (signal transduction) control over where changes occur in the genome.
   – In somatic hypermutation, only the DNA segment encoding the antigen-binding region of the antibody molecule is mutagenized. This specificity appears to be determined by specific transcription signals.

2. The specificity of immune system DNA rearrangements is compatible with flexibility in the DNA sequences produced by each rearrangement. While this is easy to understand in the case of somatic hypermutation, it also applies to V(D)J joining events. The mechanism of rejoining broken V, D and J coding segments is such that many different nucleotide sequences can occur at each novel

junction. In the process of V–D and D–J joining, the lymphocyte can even add extra untemplated nucleotides using the enzyme terminal transferase. These additional sources of coding sequence flexibility raise combinatorial diversity from the $\sim 3 \times 10^6$ different molecules achieved by accessing available V, D and J segments to over 10^{12} different molecules, providing the magnitude of diversity needed to recognize all possible foreign antigens. While the rearrangement specificity dictated by RSSs ensures that changes are limited to the antigen-binding region of the antibody molecule, joining flexibility permits the formation of an unlimited variety of binding specificities. This is how lymphocytes generate non-random diversity.

3. The computational and signaling networks that control lymphocyte differentiation also control the highly stereotyped sequence of different DNA engineering steps: V–D joining preceeds D–J joining of immunoglobulin heavy chain exons followed by V–J joining of light chain exons. An as yet undiscovered intracellular mechanism also limits productive VDJ and VJ joining to a single homologue of each chromosome pair carrying immunoglobulin coding sequences (Mostoslavsky et al., 2004). The resulting "virgin" B cell carries IgM molecules on its surface and can be stimulated to proliferate when a surface IgM molecule binds antigen. Only after the B cell undergoes mitogenic stimulation and clonal expansion does somatic hypermutation become active. Further signaling by T cell-derived lymphokines is needed for class switch recombination to differentiate the already amplified and hypermutated B cells into producers of IgG, IgE and IgA antibodies.

The immune system exemplifies two critical features of natural genetic engineering: (i) tight control lasting many cell generations over the activity of DNA rearrangement functions and (ii) targeting mechanisms to localize changes to appropriate sites in the genome. Like the ciliate example, lymphocytes also demonstrate cellular capacity to coordinate different modes of DNA restructuring to achieve a functional objective.

6. Cellular capacities for natural genetic engineering at evolutionary time scales

Even before we had the ability to sequence DNA, it was clear that natural genetic engineering is critical to evolution. The most widespread

evolutionary episode documented by direct scientific observation re-sulted from the massive application of antibiotics in medicine and agriculture following World War II. As a consequence, antibiotic-resistant strains emerged in numerous bacterial species. When antibiotic application began, we had a robust -- *and experimentally confirmed –* theory of how resistance would evolve: chromosomal mutations would cause gradual changes in cell structure that rendered the bacteria impermeable or the cellular targets insensitive to each antibiotic (Hayes, 1968). Such mutations could easily be observed in the laboratory. Nonetheless, when the molecular basis of natural antibiotic resistance was analyzed, it turned out that existing theory was completely incorrect. Resistance was due to the acquisition of new biochemical functions inactivating or removing antibiotics from the cell, and these functions were encoded by a series of DNA mobility systems: transmissible plasmids (Watanabe, 1963), transposons (Bukhari et al., 1977), and integrons (Stokes and Hall, 1989). The role of natural genetic engineering in the most extensively investigated case of contemporary evolution was impossible to ignore.

Whole genome sequencing has revealed a host of genomic features in all organisms that can only be products of natural genetic engineering. Repeats fall into this category because they cannot arise by gradual accumulation of small changes. Repeats found in genomes include the virtually ubiquitous occurrences of paralogue families (multiple copies of related protein coding sequences in a single genome; Jordan et al., 2001), segmental duplications (Arabidopsis Genome Initiative, 2000; Eichler, 2001), and the appearance of related domains in many different proteins (International Human Genome Consortium, 2001). There are various mechanisms by which duplications can arise through MGE, DSB repair and/or homologous recombination functions (Figures 1 and 2; Kazazian, 2000, 2004; Bailey et al., 2003; Shapiro, 2005a). In all cases, duplications require the coordinated action of multiple proteins and DNA sequence elements. The importance of duplicated paralogues for overall system robustness has recently been documented (Kafri et al, 2005).

It is important to recognize that current knowledge requires a hierarchical systems view of genetic determinants at all levels. Since proteins are viewed as composites of interchangeable structural and functional domains (Doolittle, 1995), each coding sequence is, in essence, a modular system composed of smaller components. Each individual genetic locus comprises coding sequences, transcriptional signals (often quite

complex in structure), and post-transcriptional processing information. The signal complexes in individual loci can confer both a high degree of specificity (*e.g.* developmental timing; Yuh et al., 2000; Arnone and Davidson, 1997) and also flexibility (*e.g.* alternative splicing; Black, 2003). By sharing common transcriptional or RNA processing signals, dispersed groups of loci can be integrated into coregulated multifunctional suites, and genomic analysis provides accumulating evidence that MGEs have played a key role in establishing these suites (Britten, 1996; Brosius, 1999; Jordan et al., 2003; Peaston et al., 2004; Shapiro, 2005b).

Molecular genetic studies and genomics have also informed us about the existence in genomes of higher-level complexes (van Driel et al., 2003). These complexes comprise more than one genetic locus under coordinated regulation, such as the mammalian globin determinants (de Laat and Grosfeld, 2003) and the Hox complexes which organize multicellular morphogenesis along different body axes (Patel and Prince, 2000; Wagner et al., 2003). It is well known that such higher-level functional complexes are subject to amplification and modification for increased functional specialization as evolution proceeds (Peterson and Davidson, 2000; Prince and Patel, 2000; de Laat and Grosfeld, 2003). Not only does the mechanism of amplification require natural genetic engineering, but the principle of taking an established routine and re-using it (sometimes in modified form) for similar tasks is basic to human engineering practice as well as to evolution. Each functional adaptation does not have to be reinvented anew.

Sometimes clusters of co-regulated determinants extend megabases in length and are visualized genomically as "syntenic" regions in which the chromosomes of related species share extended segments displaying a common order of multiple genetic loci (Mouse Genome Sequencing Consortium, 2002; Zdobnov et al., 2002; Eichler and Sankoff, 2003). For these larger complexes, much of coordinate regulation appears to occur at the level of chromatin formatting over extensive domains (van Driel et al., 2003). We know from both classical and modern studies that repetitive elements, including MGEs, play important roles in nucleating and delineating extensive chromatin domains (Spofford, 1976; Gerasimova et al., 2000; Schotta et al., 2003; Schramke and Allshire, 2003; Lippman et al., 2004; Shapiro and Sternberg, 2005).

The picture that emerges from current genomics is one where MGEs and other repeats are essential both to restructuring and to formatting the genome system architectures of different taxonomic groups. Another genome characteristic, overall size, appears to be an

important determinant of cell-cycle and organismal life-cycle duration (Cavalier-Smith, 1985; Jakob et al., 2004), and whole genome sequencing indicates that expansion occurs chiefly by the accumulation of MGEs in dispersed and tandem arrays (Kentner et al., 2003; Ma and Bennetzen, 2004; Zhang and Wessler, 2004). In agreement with this view of MGEs as central actors in evolutionary diversification is the fact that repetitive DNA is a far better indicator of taxonomic specificity than are coding sequences (summarized in Sternberg and Shapiro, 2004).

7. Impact of cellular computation on activation of natural genetic engineering functions (temporal control)

A basic aspect of computational control over natural genetic engineering functions is the ability to turn them on and off in response to appropriate inputs. There is abundant evidence that MGEs and other agents of genome restructuring are responsive to a wide variety of biological challenges, a point that Barbara McClintock emphasized in her Nobel Prize address (McClintock, 1984).

One of the most thoroughly studied situations is activation of DNA transposable elements and point mutations in bacteria by starvation stress, a phenomenon called "adaptive mutation" (Shapiro, 1997; Rosenberg, 2001). Although many authors try to limit the discussion of this phenomenon to frameshifts and base substitutions, the fact is that it was first documented with coding sequence fusions (domain joining) mediated by a transposable element (Shapiro, 1984), and there are several examples in different genera of MGE activation. In the coding sequence fusion case in *E. coli*, oxidative starvation stress increases the frequency of rearrangements by over 5 orders of magnitude (Shapiro, 1984; Maenhaut-Michel and Shapiro, 1994) and involves multiple signal transduction networks (Shapiro, 1997; Lamrani et al., 1999). The same complexity of regulation is true of transposon activation in *Pseudomonas putida* (Ilves et al., 2001) and of frameshifts in *E. coli* (Rosenberg, 2001; Lombardo et al., 2004). In yeast, adaptive mutation also involves signal transduction networks (Storchova et al., 1998; Storchova and Vondrejs, 1999) and includes the non-homologous end-joining (NHEJ) system central to many eukaryotic DNA rearrangements (Heidenreich and Eisler, 2004).

The first indications that MGEs are subject to stress activation were in maize, where McClintock found that repeated cycles of chromosome breakage led to the recovery of function by previously silenced DNA transposons (McClintock, 1987). She called the challenges that turn on MGEs "genome shock." Extensive work by her successors in maize and other plant systems has documented a wide array of stress conditions that activate both DNA- and RNA-based MGEs (Dellaporta et al., 1984; Wessler, 1996; Grandbastien, 1998; Hashida et al., 2003; Kovalchuk et al., 2003; Arnholdt-Schmitt, 2004; Madlung and Comai, 2004). In addition to chromosome breakage, activating conditions include changes in ploidy, hybridization between different species, temperature change, desiccation, high salt, tissue culture, and infection by fungal and bacterial pathogens. An important case of environmental differences in retrotransposon activity related to changes in genome size has been reported in wild barley (Kalender et al., 2000). The involvement of specific signaling networks in activation is clear because different tobacco retrotransposons respond to distinct stress signals (Beguiristain et al., 2001). Moreover, it has recently been demonstrated that genome destabilizing signals can be transmitted systemically through a tobacco plant by damage signals from a single grafted leaf (Filkowski et al., 2004).

McClintock observed in her studies that active MGEs are generally silenced within a few generations. Molecular studies in plants have documented important roles for methylation, heterochromatinization and RNA interference (RNAi) in the silencing of newly introduced or newly activated MGEs (Okamoto and Hirochika, 2001; Hashida et al., 2003). In fact, plant studies were among the first to identify the role of RNAi in silencing and led to the idea that MGE regulation is linked to heterochromatin nucleation and the establishment of epigenetic inheritance patterns (Matzke et al., 1999, 2001). Genetic studies in *Arabidopsis thaliana* have shown that various MGEs respond in characteristic ways to mutations affecting methylation, heterochromatin structure and RNAi (Lippman et al., 2003). Plant geneticists have been among the most advanced in thinking of the relationship between stress and the reprogramming of cellular computation networks maintaining genome stability (Arnholdt-Schmitt, 2004; Madlung and Comai, 2004).

In animals, studies of MGE activation have focused on the phenomenon known as "hybrid dysgenesis" (literally, abnormal reproduction in hybrids). Hybrid dysgenesis occurs when normal mating

patterns are disrupted and sperm carrying active MGEs fertilize an egg lacking them. The MGEs are latent in their original populations but become highly active in the developing germ line of a hybrid embryo, where they cause chromosome rearrangements, mutations, and insertions into new locations (Bregliano and Kidwell, 1983; Engels, 1989). Initially described in *Drosophila melanogaster*, hybrid dysgenesis has also been observed in mammals (O'Neill et al., 1998; Vrana et al., 2000). The most interesting aspect of hybrid dysgenesis is that a number of transposition and chromosome rearrangements events occur during the pre-meiotic or mitotic development of the germ line. At meiosis, clones of germ cells produce clusters of gametes that carry a constellation of multiple (non-independent) genome alterations (Woodruff and Thompson, 2002). These gametes can then participate in several fertilization events to generate a small (potentially interbreeding) population of progeny sharing complex genome rearrangements. By virtue of MGE activation during mitotic development, multiple genomic changes in sexually reproducing organisms do not have to be independent of each other nor do they have to occur in single individuals. In other words, there is a reasonable probability for the survival and proliferation of major genome reorganizations.

8. Impact of cellular computation on genome localization of natural genetic engineering activities

Contrary to the widespread assertion that cells cannot determine the sites of genetic change, there is a growing body of evidence that targeting can occur by a variety of well-defined molecular mechanisms (Shapiro, 2005a). We have already discussed some of these in looking at lymphocyte rearrangements and somatic hypermutation. Once we recognize that natural genetic engineering involves the action of nucleoprotein complexes on DNA, there is no reason to think that these complexes cannot interact with other molecular systems capable of recognizing specific DNA sequences or being directed to particular regions of the genome. There is common agreement that cellular computing networks can target transcriptional and chromatin formatting complexes to particular regions in the genome in response to the appropriate inputs. Otherwise, regulation of metabolism, the cell cycle, cellular differentiation and multicellular morphogenesis would not be possible.

The best way to understand how natural genetic engineering activities can be targeted is to consider the four mechanisms documented to date. It is almost certain that other mechanisms will be discovered in the years to come because there are cases where specificity has been documented but molecular details are unknown (Shapiro, 2005a).

1. *Sequence recognition by proteins.* Just as transcription and replication proteins can recognize specific DNA sequences, so do the proteins involved in genome restructuring. This is the basis for the specificity of DNA breakage in transposition events and in the mechanistically related case of V(D)J joining reactions. The important point is that only part of the system has to be specific. Specific sequence recognition can be combined with other DNA restructuring biochemical events that may be non-specific. This, in fact, happens in V(D)J joining when the non-specific NHEJ complex joins together exons broken at the RSS targets (Gellert, 2000; Bassing et al., 2002). Similarly, in budding yeast, the sequence-specific HO endonuclease targets non-specific homologous recombination functions to replace regulatory cassettes at the expressed mating-type (MAT) locus (Haber, 1998). Another comparable situation occurs in arthropods, where a sequence-specific endonuclease targets non-specific reverse transcription functions to insert the R1 and R2 LINE elements into particular sites in the spacer region of 28S ribosomal RNA coding repeats (Burke et al., 1999).

2. *Sequence recognition by RNA.* We are beginning to learn how important DNA recognition by RNA molecules is in transcriptional regulation and heterochromatin formatting. We know from two bacterial examples that RNA can also direct DNA restructuring to complementary sequences. One case involves the well-studied RNA-guided reverse splicing of a *Lactobacillus lactis* intron back into its target exon as a cDNA (Mohr et al., 2000). This targeting system has already been modified for use as an *in vivo* genetic engineering tool (Kahrberg et al., 2001). The other case was recently discovered in a bacterial virus that uses reverse transcription to diversify the structure of a protein that recognizes surface receptor molecules on strains of *Bordetella bronchiseptica* and *B. pertussis* (Liu et al., 2002). A diversity generating RNA targets the coding sequence by complementarity to a 21 mucleotide segment and then replaces the rest by directing reverse transcription and cDNA insertion. This appears to be a general strategy for pro-

tein diversification in bacteria because homologies to the reverse transcriptase and other components of the diversity-generating retroelement (DGR) can be recognized in a variety of prokaryotic genomes (Doulatov et al., 2004).

3. *Connections to transcriptional apparatus.* Natural genetic engineering activities and the transcriptional apparatus can interact in at least three different ways. The example we have already seen at work in immunoglobulin class switch recombination (CSR) depends on transcription rendering sites susceptible to cleavage and subsequent NHEJ joining (Kinoshito and Honjo, 2001; Chaudhuri and Alt, 2004). A different mechanism involves complex formation between a PolIII transcription initiation factor and the integrase of the yeast Ty3 LTR retrotransposon (Yieh et al., 2000). This second mechanism applies to all promoters that use that particular transcription factor and appears to place the inserted retrotransposon at a tightly defined distance from the start of transcription (Kim et al., 1998). Analogous preferences for PolII promoters are seen with various yeast retrotransposons and Murine Leukemia Virus (MLV) insertions into the human genome (Wu et al., 2003). A third mechanism involves regional targeting of genetically engineered P factor DNA transposons in *Drosophila*. If a binding site for a particular transcription factor is included in the P factor construct, it acquires a new insertion specificity and displays a high degree of "homing" to chromosome regions where that factor controls one or more genetic loci (Hama et al., 1990; Kassis et al., 1992; Fauvarque and Dura, 1993; Taillebourg and Dura, 1999; Bender and Hudson, 2000). P factor homing is particularly interesting because the insertions into a particular region have a rather broad distribution over thousands of base pairs. This means that MGEs can be targeted to suites of coregulated functions and still have the flexibility to modify those functions in different ways, such as altering transcriptional regulation, chromatin formation, or coding sequence content.

4. *Connections to chromatin formatting.* A notable feature of many MGEs is their tendency to accumulate in heterochromatic regions of the genome (e.g. Peterson-Burch et al., 2004). This phenomenon has been analyzed in the case of the Ty5 LTR retrotransposon in budding yeast. In that case, insertion is regional and is mediated by interactions between an identified targeting domain (TD) of the integrase protein and the SIR4 heterochromatin binding protein (Zou et al., 1996; Xie et al., 2001). Connecting TD to other DNA

binding proteins has been used to engineer the insertion specificity
of the Ty5 element (Zhu et al., 2003).

9. Discussion: A 21st Century/computational view of genome evolution and its consequences

The many molecular genetic details summarized and discussed above
lead to a set of basic genomic and evolutionary principles best incor-
porated into a computational perspective (Shapiro, 2002a, 2005a;
Shapiro and Sternberg, 2005):

– Genomes are formatted by repetitive elements and organized
  hierarchically for multiple information storage and transmission
  functions;
– Major evolutionary steps occur by DNA rearrangements carried
  out by diverse natural genetic engineering systems operating
  non-randomly;
– Significant evolutionary changes results from altering repetitive
  elements formatting genome system architecture as well as from
  altering protein and RNA coding sequences;
– Evolutionary changes are responsive to cellular computing net-
  works with respect to timing and location of DNA rearrange-
  ments.

These principles are computational in two complementary ways.
The first is that evolutionary change by activation of natural genetic
engineering results from cellular computation and decision-making in
crisis situations. This aspect of evolutionary computation is well doc-
umented and ever more widely accepted in the genomics community.
The second way is that natural genetic engineering can reprogram
stored genomic routines as an integral part of the computational
response to crisis. In other words, the ability to evaluate challenges
and target natural genetic engineering functions can be integrated to
optimize the chance for a successful response.

The second optimizing aspect of evolutionary computation is not
documented and remains intensely controversial. It is not difficult to
imagine the ways that computational targeting of genome restructur-
ing can optimize evolutionary efficiency by reutilizing existing func-
tional modules (domains, proteins, promoter-enhancer combinations)
and by targeting functional suites that already operate coordinately

(e.g. Hox complexes, syntenic complexes, signaling pathways, distributed suites of coregulated loci; Duboule and Wilkins, 1998; Ray et al., 2004, and other articles the same issue of Science). It is also rather easy to see how the mobilization of existing genomic subsystems can generate the internally duplicated and hierarchic genome system architectures that we continue to discover.

The computational perspective outlined above presents challenges to two groups of scientists interested in developing a deeper, more modern view of evolution. For the computer scientists, it asks that they convert the molecular interaction data into a manageable set of formalisms and computer simulations that will allow us to investigate analytically how effective known natural genetic engineering processes are in promoting evolutionary novelties. For the molecular geneticists, it asks that we develop more sophisticated experimental protocols to investigate the generation of complex genome circuits and test the importance of targeting mechanisms in their origins. These challenges should be high on the research agenda for the 21st Century. It is likely that meeting them will lead us to new computing paradigms of great creative power, like evolution itself.

References

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

Arnholdt-Schmitt B (2004) Stress-induced cell reprogramming. A role for global genome regulation?. Plant Physiology 136: 2579–2586

Arnone MI and Davidson EH (1997) The hardwiring of development: organisation and function of genomic regulatory systems. Development 124: 1851–1864

Bailey JA, Liu G and Eichler EE (2003) An Alu transposition model for the origin and expansion of human segmental duplications. American Journal of Human Genetics 73: 823–834

Bailey JA, Baertsch R, Kent WJ, Haussler D and Eichler EE (2004) Hotspots of mammalian chromosomal evolution. Genome Biology 5: R23. Epub 2004 Mar 08

Bassing CH, Swat W and Alt FW (2002) The mechanism and regulation of chromosomal V(D)J recombination. Cell 109: S45–S55

Bender W and Hudson A (2000) P element homing to the *Drosophila* bithorax complex. Development 127: 3981–3992

Beguiristain T, Grandbastien M-A, Puigdomenech P and Casacuberta JM (2001) Three Tnt1 subfamilies show different stress-associated patterns of expression in tobacco. Consequences for retrotransposon control and evolution in plants. Plant Physiology 127: 212–221

Black DL (2003) Mechanisms of alternative pre-messenger RNA splicing. Annual Reviews of Biochemistry 72: 291–336

Bolouri H and Davidson EH (2002) Modeling transcriptional regulatory networks. Bioessays 24: 1118–29

Bray D (1990) Intracellular signalling as a parallel distributed process. Journal of Theoretical Biology 143: 215–231

Bray D (1995) Protein molecules as computational elements in living cells. Nature 376: 307–312

Bregliano JC and Kidwell M (1983) Hybrid Dysgenesis. In: Shapiro JA (ed.) Mobile Genetic Elements, pp. 363–410. Academic Press, NY

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

Brookfield JF (2004) Evolutionary genetics: Mobile DNAs as sources of adaptive change?. Current Biology 14: R344–345

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

Bukhari AI, Shapiro JA and Adhya SL (1977) .DNA Insertion Elements, Episomes and Plasmids. Cold Spring Harbor Press, Cold Spring Harbor, NY

Burke WD, Malik HS, Jones JP and Eickbush TH (1999) The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. Molecular Biology and Evolution 16: 502–511

Burnet M (1964) A Darwinian approach to immunity. Nature 203: 451–454

Cavalier-Smith T (1985) The Evolution of Genome Size. John Wiley & Sons Ltd., Chichester

Chaudhuri J and Alt FW (2004) Class-switch recombination: interplay of transcription, DNA deamination and DNA repair. Nature Reviews of Immunology 4: 541–552

Craig NL, Craigie R, Gellert M and Lambowitz AM (2002) .Mobile DNA II. ASM Press, Washington DC

de Laat W and Grosveld F (2003) Spatial organization of gene expression: The active chromatin hub. Chromosome Research 11: 447–459

Dellaporta SL, Chomet PS, Mottinger JP, Wood JA, Yu SM and Hicks JB (1984) Endogenous transposable elements associated with virus infection in maize. Cold Spring Harbor Symposium of Quantitative Biology 49: 321–328

Doolittle RF (1995) The multiplicity of domains in proteins. Annual Review of Biochemistry 64: 287–314

Doulatov S, Hodes A, Dai L, Mandhana N, Liu M, Deora R, Simons RW, Zimmerly S and Miller JF (2004) Tropism switching in *Bordetella* bacteriophage defines a family of diversity-generating retroelements. Nature 431: 476–481

Duboule D and Wilkins AS (1998) The evolution of 'bricolage'. Trends in Genetics 14: 54–59

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

Eichler EE and Sankoff D (2003) Structural dynamics of eukaryotic chromosome evolution. Science 301: 793–797

Engels WR (1989) P elements in *Drosophila melanogaster*. In: Berg DE and Howe MM (eds) Mobile DNA, pp. 437–484. ASM Press, Washington, DC

Fauvarque MO and Dura JM (1993) Polyhomeotic regulatory sequences induce developmental regulator-dependent variegation and targeted P-element insertions in *Drosophila*. Genes and Development 7: 1508–1520

Filkowski J, Yeoman A, Kovalchuk O and Kovalchuk I (2004) Systemic plant signal triggers genome instability. Plant Journal 38: 1–11

Franklin A and Blanden RV (2004) On the molecular mechanism of somatic hypermutation of rearranged immunoglobulin genes. Immunology and Cell Biology 82: 557–567

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

Gerasimova TI, Byrd K and Corces VG (2000) A chromatin insulator determines the nuclear localizations of DNA. Molecular Cell 6: 1025–1035

Goday C and Esteban MR (2001) Chromosome elimination in sciarid flies. Bioessays 23: 242–50

Grandbastien M-A (1998) Activation of plant retrotransposons under stress conditions. Trends in Plant Science 3: 181–187

Greider CW and Blackburn EH (1987) The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. Cell 51: 887–898

Haber JE (1998) Mating-type gene switching in Saccharomyces cerevisiae. Annual Reviews of Genetics 32: 561–599

Hall BG (1988) Adaptive evolution that requires multiple spontaneous mutations. I. Mutations involving an insertion sequence. Genetics 120: 887–897

Hama C, Ali Z and Kornberg TB (1990) Region-specific recombination and expression are directed by portions of the Drosophila engrailed promoter. Genes and Development 4: 1079–1093

Harden N and Ashburner M (1990) Characterization of the FB-NOF transposable element of Drosophila melanogaster. Genetics 126: 387–400

Hartwell LH, Hopfield JJ, Leibler S and Murray AW (1999) From molecular to modular cell biology. Nature 402: C47–52

Hashida SN, Kitamura K, Mikami T and Kishima Y (2003) Temperature shift coordinately changes the activity and the methylation state of transposon Tam3 in Antirrhinum majus. Plant Physiology 132: 1207–1216

Hayes W (1968) The Genetics of Bacteria and their Viruses. 2nd edition, Blackwell, London

Heidenreich E and Eisler H (2004) Non-homologous end joining dependency of gamma-irradiation-induced adaptive frameshift mutation formation in cell cycle-arrested yeast cells. Mutation Research 556: 201–208

Ilves H, Horak R and Kivisaar M (2001) Involvement of sigma(S) in starvation-induced transposition of Pseudomonas putida transposon Tn4652. Journal of Bacteriology 183: 5445–5448

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

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

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

Jahn CL and Klobutcher LA (2002) Genome remodeling in ciliated protozoa. Annual Reviews of Microbiology 56: 489–520

Jakob SS, Meister A and Blattner FR (2004) The considerable genome size variation of *Hordeum* species (*Poaceae*) is linked to phylogeny, life form, ecology, and speciation rates. Molecular Biology and Evolution 21: 860–869

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

Jessop-Murray H, Martin LD, Gilley D, Preer JJ and Polisky B (1991) Permanent rescue of a non-Mendelian mutation of *Paramecium* by microinjection of specific DNA sequences. Genetics 129: 727–34

Jordan IK, Makarova KS, Spouge JL, Wolf YI and Koonin EV (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. Genome Research 11: 555–565

Jordan IK, Rogozin IB, Glazko GV and Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends in Genetics 19: 68–72

Kafri R, Bar-Even A and Pilpel Y (2005) Transcription control reprogramming in genetic backup circuits. Nature genetics 37: 295–299

Kalendar R, Tanskanen J, Immonen S, Nevo E and Schulman A (2000) Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proceedings of the National Academy of Sciences USA 97: 6603–6607

Karberg M, Guo H, Zhong J, Coon R, Perutka J and Lambowitz AM (2001) Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. Nature Biotechnology 19: 1162–1167

Kassis JA, Noll E, Vansickle EP, Odenwald WF and Perrimon N (1992) Altering the insertional specificity of a *Drosophila* transposable element. Proceedings of the National Academy of Sciences USA 89: 1919–1923

Kazazian HH Jr (2000) L1 retrotransposons shape the mammalian genome. Science 289: 1152–1153

Kazazian HH Jr (2004) Mobile elements: Drivers of genome evolution. Science 303: 1626–1632

Kentner EK, Arnold ML and Wessler SR (2003) Characterization of high-copy-number retrotransposons from the large genomes of the Louisiana iris Species and their use as molecular markers. Genetics 164: 685–697

Kim JM, Vanguri S, Boeke JD, Gabriel A and Voytas DF (1998) Transposable elements and genome prganization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Research 8: 464–478

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

Klobutcher LA and Herrick G (1997) Developmental genome reorganization in ciliated protozoa: The transposon link. Progress in Nucleic Acid Research and Molecular Biology 56: 1–62

Kovalchuk I, Kovalchuk O, Kalck V, Boyko V, Filkowski J, Heinlein M and Hohn B (2003) Pathogen induced systemic plant signal triggers DNA rearrangements. Nature 423: 760–762

Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE and Cech TR (1982) Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. Cell 31: 147–57

Lamrani S, Ranquet C, Gama M-J, Nakai H, Shapiro JA, Toussaint A and Maenhaut-Michel G (1999) Starvation-induced Mucts62-mediated coding sequence fusion: roles for ClpXP, Lon, RpoS and Crp. Molecular Microbiology 32: 327–343

Landweber LF and Kari L (1999) The evolution of cellular computing: nature's solution to a computational problem. Biosystems 52: 3–13

Lippman Z, May B, Yordan C, Singer T and Martienssen R (2003) Distinct mechanisms determine transposon inheritance and methylation via small interfering RNA and histone modification. PLoS Biology 1: E67

Lippman Z, Gendrel AV, Black M, Vaughn MW and Dedhia N, et al., (2004) Role of transposable elements in heterochromatin and epigenetic control. Nature 430: 471–476

Liu M, Deora R, Doulatov SR, Gingery M, Eiserling FA, Preston A, Maskell DJ, Simons RW, Cotter PA, Parkhill J and Miller JF (2002) Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage. Science 295: 2091–4

Lombardo MJ, Aponyi I and Rosenberg SM (2004) General stress response regulator RpoS in adaptive mutation and amplification in Escherichia coli. Genetics 166: 669–680

Ma J and Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. Proceedings of the National Academy of Sciences USA 101: 12404–12410

Madlung A and Comai L (2004) The effect of stress on genome regulation and structure. Annals of Botany 94: 481–495

Maenhaut-Michel G and Shapiro JA (1994) The roles of starvation and selective substrates in the emergence of araB-lacZ fusion clones. EMBO Journal 13: 5229–5239

Matzke MA, Mette MF, Aufsatz W, Jakowitsch J and Matzke AJ (1999) Host defenses to parasitic sequences and the evolution of epigenetic control mechanisms. Genetica 107: 271–287

Matzke M, Matzke AJ and Kooter JM (2001) RNA: Guiding gene silencing. Science 293: 1080–1083

McClintock B (1984) Significance of responses of the genome to challenge. Science 226: 792–801

McClintock B (1987) Discovery And Characterization of Transposable Elements: The Collected Papers of Barbara McClintock. Garland, New York, NY

Meyer E and Garnier O (2002) Non-Mendelian inheritance and homology-dependent effects in ciliates. Advances in Genetics 46: 305–338

Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, Sheffer M and Alon U (2004) Superfamilies of evolved and designed networks. Science 303: 1538–1542

Mohr G, Smith D, Belfort M and Lambowitz AM (2000) Rules for DNA target-site recognition by a lactococcal group II intron enable retargeting of the intron to specific DNA sequences. Genes and Development 14: 559–573

Moran JV, DeBerardinis RJ and Kazazian HH Jr (1999) Exon shuffling by L1 retrotransposition. Science 283: 1530–1534

Mostoslavsky R, Alt FW and Bassing CH (2003) Chromatin dynamics and locus accessibility in the immune system. Nature Immunology 4: 603–606

Mostoslavsky R, Alt FW and Rajewsky K (2004) The lingering enigma of the allelic exclusion mechanism. Cell 118: 539–544

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

Muller F and Tobler H (2000) Chromatin diminution in the parasitic nematodes *Ascaris suum* and *Parascaris univalens*. International Journal of Parasitology 30: 391–399

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

Okamoto H and Hirochika H (2001) Silencing of transposable elements in plants. Trends in Plant Science 6: 527–534

O'Neill RJ, O'Neill MJ and Graves JA (1998) Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. Nature 393: 68–72

Patel NH and Prince VE (2000) Beyond the Hox complex. Genome Biol. 1, reviews 1027.1–1027.4. The electronic version of this article is the complete one: http://genomebiology.com/2000/1/5/reviews/1027

Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D and Knowles BB (2004) Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Developmental Cell 7: 597–606

Peterson KJ and Davidson EH (2000) Regulatory evolution and the origin of the bilaterians. Proceedings of the National Academy of Sciences USA 97: 4430–4433

Peterson-Burch BD, Nettleton D and Voytas DF (2004) Genomic neighborhoods for *Arabidopsis* retrotransposons: a role for targeted integration in the distribution of the Metaviridae. Genome Biology 5: R78

Prescott DM (2000) Genome gymnastics: Unique modes of DNA evolution and processing in ciliates. Nature Reviews in Genetics 1: 191–198

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

Ray LB, Adler EM and Gough NR (2004) Common signaling themes. Science 306: 1505

Redi CA, Garagna S, Zacharias H, Zuccotti M and Capanna E (2001) The other chromatin. Chromosoma 110: 136–147

Rosenberg SM (2001) Evolving responsively: Adaptive mutation. Nature Reviews: Genetics 2: 504–15

Ruben AJ and Landweber LF (2000) The past, present and future of molecular computing. Nature Reviews in Molecular and Cell Biology 1: 69–72

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

Schramke V and Allshire R (2003) Hairpin RNAs and retrotransposon LTRs effect RNAi and chromatin-based gene silencing. Science 301: 1069–1074

Shapiro JA (1979) A molecular model for the transposition and replication of bacteriophage Mu and other transposable elements. Proceedings of the National Academy of Sciences USA 76: 1933–1937

Shapiro JA. (1983) .Mobile Genetic Elements. Academic Press, New York

Shapiro JA (1984) Observations on the formation of clones containing *araB-lacZ* cistron fusions. Molecular and General Genetics 194: 79–90

Shapiro JA (1992) Natural genetic engineering in evolution. Genetica 86: 99–111

Shapiro JA (1997) Genome organization, natural genetic engineering, and adaptive mutation. Trends in Genetics 13: 98–104

Shapiro JA (1999) Genome system architecture and natural genetic engineering in evolution. In: Caporale L (ed) Molecular Strategies for Biological Evolution. Annals of the NY Academy of Science 870: 23–35

Shapiro JA (2002a) Genome organization and reorganization in evolution: Formatting for computation and function. In: Speybroeck V, de Vijver GV and de Waele D (eds) From Epigenesis to Epigenetics: The Genome in Context. Annals of the NY Academy of Science 981: 111–134

Shapiro JA (2002b) A 21st Century view of evolution. Journal of Biological Physics 28: 745–764

Shapiro JA (2005a) A 21st Century view of evolution: Genome system architecture, repetitive DNA, And natural genetic engineering. Gene 345: 91–100

Shapiro JA (2005b) Retrotransposons and regulatory suites. BioEssays 27: 122–125

Shapiro JA, Bukhari AI and Adhya SL (1977) New Pathways in the evolution of chromosome structure. In: Bukhari AI et al. (eds) DNA Insertion Elements, Episomes and Plasmids, pp. 3–13. Cold Spring Harbor Press, Cold Spring Harbor NY

Shapiro JA and Sternberg RV (2005) Why repetitive DNA is essential for genome function. Biol Rev Camb Philos Soc 80: 227–250

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

Sternberg RV Shapiro JA (2005) How repeated retroelements format genome function. Cytogenet. Genome Res 110: 108–116

Stokes HW and Hall RM (1989) A novel family of potentially mobile DNA elements encoding site-specific gene-integration functions: Integrons. Molecular Microbiology 3: 1669–1683

Storchová Z, Rojas Gil AP, Janderová B and Vondrejs V (1998) The involvement of the RAD6 gene in starvation-induced reverse mutation in *Saccharomyces cerevisiae*. Molecular and General Genetics 258: 546–552

Storchova Z and Vondrejs V (1999) Starvation-associated mutagenesis in yeast *Saccharomyces cerevisiae* is affected by Ras2/cAMP signaling pathway. Mutation Research 431: 59–67

Taillebourg E and Dura JM (1999) A novel mechanism for P element homing in *Drosophila*. Proceedings of the National Academy of Sciences USA 96: 6856–6861

Driel Rvan, Fransz PF and Verschure PJ (2003) The eukaryotic genome: A system regulated at different hierarchical levels. Journal of Cell Science 116: 4067–4075

Vrana PB, Fossella JA, Matteson PG, O'Neill MJ and Tilghman SM (2000) Genetic and epigenetic incompatibilities underlie hybrid dysgenesis in *Peromyscus*. Nature Genetics 25: 120–124

Wagner GP, Amemiya C and Ruddle F (2003) Hox cluster duplications and the opportunity for evolutionary novelties. Proceedings of the National Academy of Sciences USA 100: 14603–14606

Watanabe T (1963) Infective heredity of multiple drug resistance in bacteria. Bacteriological Reviews 27: 87–115

Wessler SR (1996) Turned on by stress: Plant retrotransposons. Current Biology 6: 959–961

Woodruff RC and Thompson JN Jr (2002) Mutation and premating isolation. Genetica 116: 371–382

Wu X, Li Y, Crise B and Burgess SM (2003) Transcription start regions in the human genome are favored targets for MLV integration. Science 300: 1749–1751

Xie W, Gai X, Zhu Y, Zappulla DC, Sternglanz R and Voytas DF (2001) Targeting of the Yeast Ty5 Retrotransposon to Silent Chromatin Is Mediated by Interactions between Integrase and Sir4p. Molecular and Cellular Biology 21: 6606–6614

Yieh L, Kassavetis G, Geiduschek EP and Sandmeyer SB (2000) The Brf and TATA-binding protein subunits of the RNA polymerase III transcription factor IIIB mediate position-specific integration of the gypsy-like element, Ty3. Journal of Biological Chemistry 275: 29800–29807

You Y, Aufderheide K, Morand J, Rodkey K and Forney J (1991) Macronuclear transformation with specific DNA fragments controls the content of the new macronuclear genome in *Paramecium tetraurelia*. Molecular and Cellular Biology 11: 1133–37

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

Zdobnov EM, Mering Cvon , Letunic I, Torrents D, Suyama M, Copley RR, Christophides GK, Thomasova D, Holt RA and Subramanian G et al. (2002) Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. Science 298: 149–159

Zhang X and Wessler SR (2004) Genome-wide comparative analysis of the transposable elements in the related species *Arabidopsis thaliana* and *Brassica oleracea*. Proceedings of the National Academy of Sciences USA 101: 5589–5594

Zhu Y, Dai J, Fuerst PG and Voytas DF (2003) Controlling integration specificity of a yeast retrotransposon. Proceedings of the National Academy of Sciences USA 100: 5891–5895

Zou S, Ke N, Kim JM and Voytas DF (1996) The *Saccharomyces* retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci. Genes and Development 10: 634–645