

Review

# A 21st century view of evolution: genome system architecture, repetitive DNA, and natural genetic engineering

James A. Shapiro

*Department of Biochemistry and Molecular Biology, University of Chicago, 920 E. 58th Street, Chicago, IL 60637, United States*

Received 4 September 2004; received in revised form 20 October 2004; accepted 9 November 2004

Available online 4 January 2005

Received by U. Bastolla

## Abstract

The last 50 years of molecular genetics have produced an abundance of new discoveries and data that make it useful to revisit some basic concepts and assumptions in our thinking about genomes and evolution. Chief among these observations are the complex modularity of genome organization, the biological ubiquity of mobile and repetitive DNA sequences, and the fundamental importance of DNA rearrangements in the evolution of sequenced genomes. This review will take a broad overview of these developments and suggest some new ways of thinking about genomes as sophisticated informatic storage systems and about evolution as a systems engineering process.

© 2004 Elsevier B.V. All rights reserved.

*Keywords:* Transposable element; Repetitive DNA; Transcriptional regulation; Chromatin domains; Biocomputing; Systems biology; Data storage

## 1. Introduction: DNA as a data storage medium

One of the keys to a 21st Century vision of how genomes operate is to think about DNA as a data storage medium that operates over three different time scales:

- Many organismal generations: genetic storage in local DNA sequences and long range chromosome structure;
- Multiple cell generations: epigenetic storage in covalent modifications and stable chromatin configurations;
- Within a single cell cycle: computational storage in meta-stable nucleoprotein complexes.

These three time scales reflect the different ways that DNA interacts with the rest of the cell as it carries out computations and decision-making. Cellular computations involve evaluation of multiple internal and external inputs. Inputs include the replication status of the genome, where the cell is in the cell cycle, what nutrients are available, what intercellular signaling molecules are present, and what other cells are touching the cell surface. Some situations require fast

responses, such as a change in the nutritional environment or the detection of genome damage. Other situations result in longer-term cellular differentiations, characterized by the formation of stable chromatin configurations (Van Driel et al., 2003). Certain conditions involve restructuring of the genome, either as part of the normal life cycle (Beermann, 1977; Prescott, 2000; Bassing et al., 2002; Kinoshita and Honjo, 2001) or in response to a crisis situation (McClintock, 1984; Shapiro, 1992, 1997).

## 2. Genome system architecture and repetitive DNA

Genomes contain several different kinds of functional information. In addition to the widely recognized coding sequences (data files) determining the primary structures of RNA and protein molecules, there is information for other essential genomic processes:

- Packaging DNA molecules within the nucleoid or nucleus;
- DNA replication and transmission of genome copies to progeny cells;

*E-mail address:* [jsha@uchicago.edu](mailto:jsha@uchicago.edu).

- Repair of DNA damage;
- DNA restructuring.

Our current understanding of how coding sequence expression (data file access) and all these other processes operate is based upon the definition of *cis*-acting signals as part of the operon and replicon theories in the early 1960s (Jacob and Monod, 1961; Jacob et al., 1963). These *cis*-acting signals are fundamentally different from any classical definition of a gene. They serve to format coding sequences and genome architecture in the same way that generic bit strings format the encoded information in electronic data storage media and guide the computational hardware to the right data files and indicate the appropriate routines to apply. *Cis*-acting signals in the genome similarly direct cellular hardware to form functional nucleoprotein complexes to carry out tasks such as transcription, replication, DNA distribution to daughter cells, and homology-dependent and homology-independent recombination (Shapiro, 2002a). Since they are generic and work at many locations, *cis*-acting signals belong to the repetitive component of the genome (Shapiro and Sternberg, 2005). By applying an informatic perspective, we can appreciate the functional relevance and interconnections of genome features which have proved difficult to understand within the linear conceptual framework of classical genetics.

Extending the informatic metaphor, it is possible to argue that genomes each have a characteristic “system architecture,” in much the same way that different computer systems do (Shapiro, 1999; Shapiro and Sternberg, 2005). The taxonomically specific system architecture includes elements such as:

- transcription signals used to regulate expression of particular coding sequences;
- signals for genome transmission (origins, centromeres and telomeres);
- signals for recombination and DNA rearrangement;
- signals for compacting the genome with protein and RNA to form particular chromatin structures;
- signals for attaching the genome to particular cellular or nuclear structures.

From the genome system architecture perspective, it is possible for two genomes in different species to have identical coding sequences but distinct signals and genome system architectures. The result of different architectures would almost certainly be germ-line reproductive incompatibility and, quite probably, distinct patterns of coding sequence expression leading to phenotypic and ecological diversity. The major determinants of genome system architecture are the repetitive elements in the genome, such as tandemly arrayed repeats at centromeres (Choo, 2001), telomere repeats that permit the replication of chromosome ends (Blackburn, 2001), and dispersed repeats that contain many signals for transcription, chromatin organization, and

nuclear localization (Jordan et al., 2003; Shapiro and Sternberg, 2005).

There is an extensive literature on the effects of repetitive DNA on coding sequence expression, including countless experiments with mobile genetic elements (Bukhari et al., 1977; Shapiro, 1983; Berg and Howe, 1989; Deininger et al., 2003) and a growing number of studies of “position effect” phenomena, where the expression of a particular genetic locus depends upon its location relative to “heterochromatic” (differently staining) blocks of repetitive elements (Spofford, 1976; Schotta et al., 2003). Particularly important are *trans*-position effects of repetitive elements on expression of genetic loci from different chromosomes (Spofford, 1976). From a mechanistic point of view, we now explain these dosage-dependent genome-wide effects as due to titration of a limited supply of chromatin-binding proteins (Schotta et al., 2003). From an organizational point of view, distant effects of repetitive element dosage tell us that the whole genome is a single integrated system, regulated both in *cis*- and *trans*- by networks employing DNA repeats.

It has been evident for a long time that repetitive DNA is a more discriminating indicator of hereditary relationships than coding sequences. For example, 25 years ago restriction site polymorphisms in tandem repeats of “alpha satellite” DNA at centromeres permitted the construction of a primate phylogeny (Donehower and Gillespie, 1979), and each mammalian order can be distinguished by its content of highly repeated SINE elements dispersed throughout the genome (generally present at between  $10^4$ – $10^6$  copies per haploid genome; data tabulated in Sternberg and Shapiro, 2005). Plant species can also be distinguished by their centromeric repeats (Shapiro and Sternberg, 2005), and closely related “sibling” *Drosophila* species differ markedly in their content of both tandem satellite arrays and dispersed repeats (Dowsett, 1983; Csink and Henikoff, 1998). Indeed, we use repetitive microsatellite DNA for forensic DNA analysis to determine relationships between individuals (Bennett, 2000). In other words, the repetitive component of the genome is far more taxonomically specific than coding sequences. This conclusion is consistent with a key role for repetitive DNA in evolutionary diversification.

### 3. Genomes and cellular computation: *E.coli lac* operon and the lessons of sequenced genomes

Informatically, we understand best how the genome interacts with the rest of the cell to carry out computations and decision-making at the shortest time scales in relatively “simple” systems, such as the classic case of the *E. coli lac* operon (Jacob and Monod, 1961). This system has been described many times from either a molecular or computational perspective (e.g. Reznikoff, 1992; Shapiro, 2002b). A series of highly integrated molecular interactions allows *E. coli* cells to distinguish between two sugars and execute the

following non-trivial algorithm: “IF lactose is available AND IF glucose is not available AND IF the cell can synthesize beta-galactosidase and lactose permease, THEN transcribe *lacZYA* from the *lac* promoter.”

To save space here, I refer the reader to other reviews for details of the *lac* operon system and its algorithmic properties. For our purposes, it is important to summarize the general conclusions one can draw from this example:

- Weak interactions, specific binding and cooperativity are essential aspects of molecular computations in cells.
- Repetition in DNA and proteins means that specific logical operations arise through combinations of basic circuit elements (e.g. complex regulatory regions in DNA, intra- and intermolecular interactions between protein domains).
- Allostery, the fact that binding of one ligand affects binding a distinct ligand, confers communication and processing capabilities on individual molecules so that cellular network nodes act as complex microprocessors.
- Layering of weak and “fuzzy” interactions provides overall sharpness to integrated cellular responses (i.e. cells operate by Fuzzy Logic principles; Zadeh, 1975).
- Cells use chemical symbols to represent physiological information.
- No separation exists between control molecules and execution molecules, telling us we cannot apply Cartesian dualism models to the *E. coli* cell, or any other cell.
- Participation of DNA directly in formation of repression and transcription nucleoprotein complexes suggests that it may also not be useful to apply Turing’s concepts of separate “machine” and “tape” (Turing, 1950) to cellular computations.

This list indicates that the principles underlying cellular analog computing may well be different from those that operate in electronic digital computers. Such a difference does not invalidate the informatic metaphor. But it does mean that we will have to be careful in applying existing computational models to cells. Combinatorics, fuzzy logic models, and principles learned from linguistics and semiotics may all serve as key guides to a formal description of cellular information-processing networks. In addition, we need to recognize that bioinformatics is far more than the application of contemporary technology to large data bases. Bioinformatics has the potential to lead us to novel computing paradigms that may prove far more powerful than the Turing machine-based digital concepts we now use. After all, no human contrivance operates with either the degree of complexity, the precision, or the efficiency of living cells.

Genome sequence analysis is one of our most important guides to disentangling how cellular systems operate and how function changes in the course of evolution. Here we find support for some of the general principles deducible from individual cases like *lac*. In particular, repetition,

reuse and combinatorics have proven to be fundamental in protein and whole genome evolution. We have realized that protein structures evolve by iterating, shuffling and accumulating domains (Doolittle, 1995; International Human Genome Consortium, 2001) and that protein families, which characterize individual taxa, evolve by coding sequence amplification (e.g. Zdobnov et al., 2002; Mouse Genome Sequencing Consortium, 2002). We see that expression systems evolve by combining coding sequences (data files), regulatory signals and chromatin markers into higher order complexes that persist and diversify in the course of evolution (e.g. homeobox domains, Patel and Prince, 2000). At even higher levels of organization, we find that genomes contain extensive chromosome segments (syntenic regions) that may be duplicated at various locations within the genome (Arabidopsis Genome Initiative, 2000; Eichler, 2001) and scrambled into new combinations during evolution (Mouse Genome Sequencing Consortium, 2002).

#### 4. Natural genetic engineering

All the preceding whole-genome sequence discoveries implicate cut-and-paste type DNA rearrangements as basic evolutionary processes. What do we know about the capacity of cells to carry out such natural genetic engineering? An important clue is the discovery that our own genomes are at least 43% composed of DNA segments that can transpose from one location to another (International Human Genome Consortium, 2001). Two classes of transposable or mobile genetic elements have been recognized from the work of Barbara McClintock and her molecular followers (McClintock, 1987; Bukhari et al., 1977; Shapiro, 1983; Berg and Howe, 1989; Craig et al., 2002; Deininger et al., 2003): DNA transposons move exclusively at the level of DNA molecules while retrotransposons and other retroelements move by means of an RNA intermediate that can be reverse-transcribed into genomic DNA (Coffin et al., 1997; Kazazian, 2000).

McClintock discovered mobile genetic elements in the first instance because they mediated chromosome rearrangements. Molecular analysis has confirmed that the same mechanisms which lead defined segments of DNA to move from one location to another (transpose) can also mediate both large- and small-scale rearrangements. There appears to be something of a molecular division of labor: DNA elements mediate rearrangements of large segments (Fig. 1), while retroelements mobilize smaller segments, generally not larger than several kilobases in length (Fig. 2). The mechanisms underlying these rearrangements are just the kind of processes needed to explain the patterns of genome conservation and scrambling found by comparing whole genome sequences. There is abundant documentation that these mechanisms have been used in evolution (Britten, 1997; Brosius, 1999; Nekrutenko and Li, 2001; Bailey et al.,

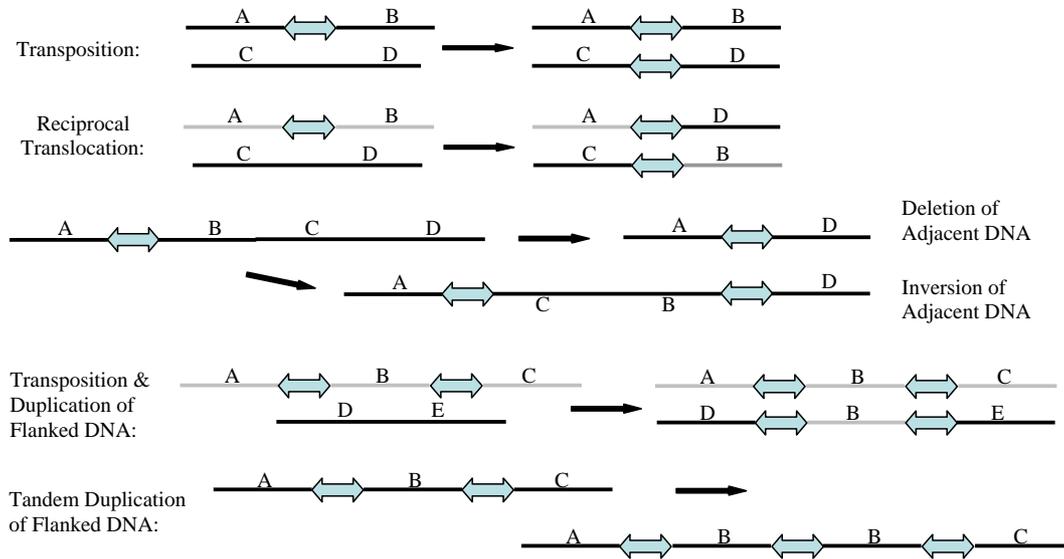


Fig. 1. Some of the rearrangements mediated by DNA transposons. The cartoons show well-documented DNA rearrangements carried out by the transposition systems of replicative or cut-and-paste DNA transposons (double-headed arrows). Often the structures of rearranged DNA generated by either mechanism are the same. These examples are based largely on the rearrangements described in Shapiro, 1979, and by Engels at <http://engels.genetics.wisc.edu/Pelements/HEI.html>. The web site includes an animation of P factor-mediated duplication/deletion events. Note that DNA transposons have at least two ways of duplicating a sequence flanked by copies of the element: either a transposed duplication at a new genetic site or a tandem duplication at the original site. Both sorts of duplications are found in sequenced genomes, especially in loci encoding large paralogue families.

2003; Jordan et al., 2003), they occur in nature (Bregliano and Kidwell, 1983; Engels, 1989; Prescott, 2000; Lerman et al., 2003), and they can execute key evolutionary processes in the laboratory, like exon shuffling (Moran et al., 1999).

An especially illuminating example of natural genetic engineering is the mammalian immune system. This system evolved from DNA transposons and cellular repair functions (Agrawal et al., 1998; Bassing et al., 2002; Gellert, 2002). It ensures the rapid evolution in lymphocytes of a virtually

infinite array of antigen-recognition protein domains starting with a finite set of germ line coding elements. If we examine how antibody heavy- and light-chain molecule-encoding DNA sequences form in B lymphocytes, we see a highly stereotypic series of DNA rearrangements with a number of instructive features (Figs. 3 and 4):

- The rearrangements occur at specific sites in the genome demarcated by a series of repetitive DNA signals,

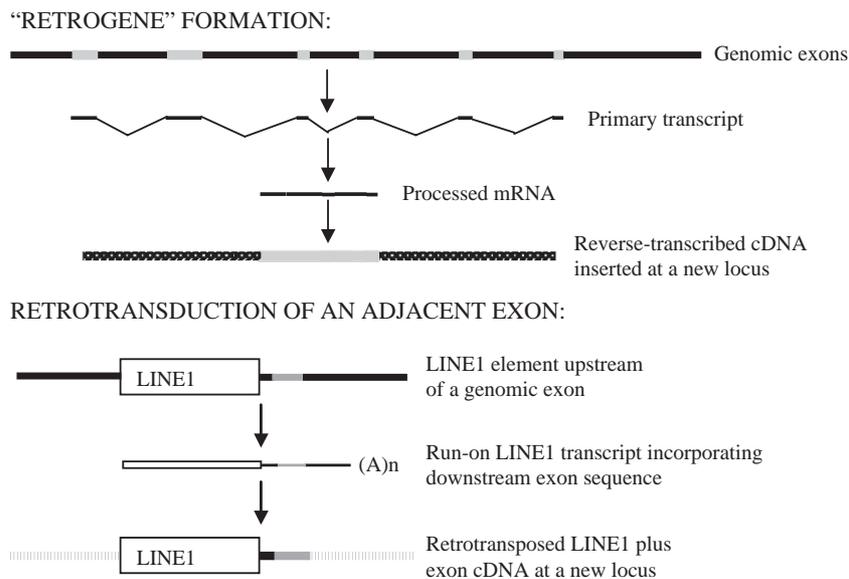


Fig. 2. Retrogene formation and retrotransduction of exons. The diagram summarizes how the reverse transcription and integration activities of LINE elements can create processed intron-free integrated cDNA copies of any cellular mRNA (“retrogenes”) or can integrate DNA copies of exons located downstream of an active LINE element after read-through transcription (“retrotransduction”). See Brosius (1999) for more details.

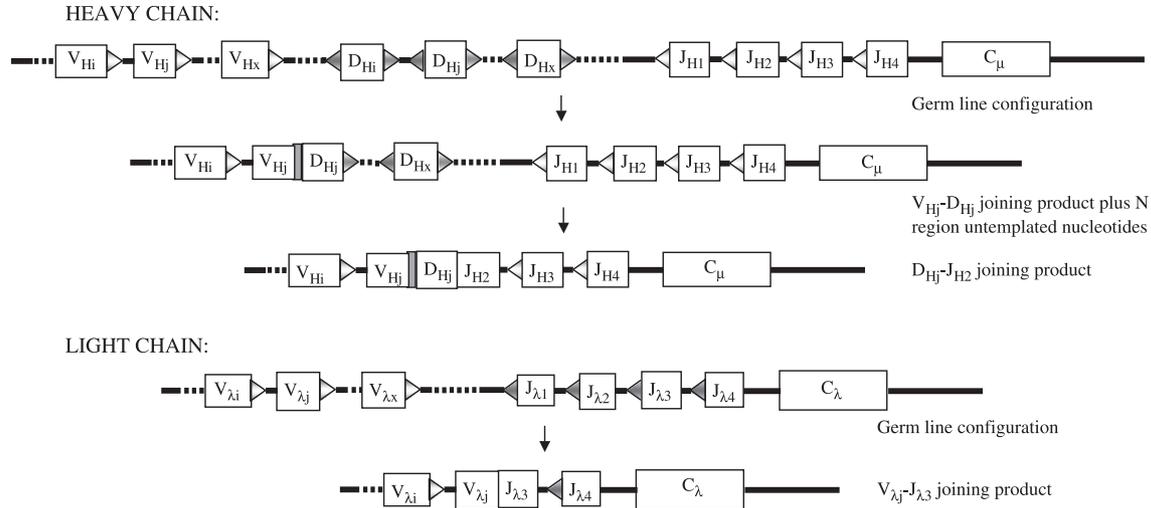


Fig. 3. Structure of immunoglobulin-coding DNA and the process of V(D)J joining. The different V, D, J and C exons and the details of the molecular events are explained in Bassing et al. (2002) and Gellert (2002). The differently shaded triangles represent complementary recombination signal sequences (RSSs). For any two exons to join together, they must be flanked by complementary RSSs. Two identical RSSs will not promote DNA breakage and rejoining. Note, in the heavy chain chromosome, how the arrangement of RSSs prevents V–J and D–D joining and effectively prevents further joining activity on the fully rearranged chromosome. The shaded rectangle in the V<sub>Hj</sub>–D<sub>Hj</sub> joining product indicates a segment of “N region” untemplated nucleotides arising from the action of terminal transferase before the broken fragments are ligated together.

- “recombination signal sequences” (RSSs) and “switch” (S) regions.
- The exact rearrangements that joint exons next to RSSs are themselves highly flexible so that junctions of variable (V), “join” (J) and “diversity” (D) exons can occur at several different internucleotide positions.
- The B cells can insert untemplated “N region” sequences next to D region sequences through the action of the enzyme terminal deoxynucleotide transferase.
- The DNA rearrangement process normally occurs only in cells destined to produce antibodies and follows a highly determined sequence (V–D joining, then D–J joining, then V–J joining, and finally class switch rearrangements [CSR] at S regions in antibody-producing cells).
- The site of DNA rearrangement is sensitive to both internal feedback (“allelic exclusion”) and external stimuli (lymphokine-directed choice of S regions for CSR).

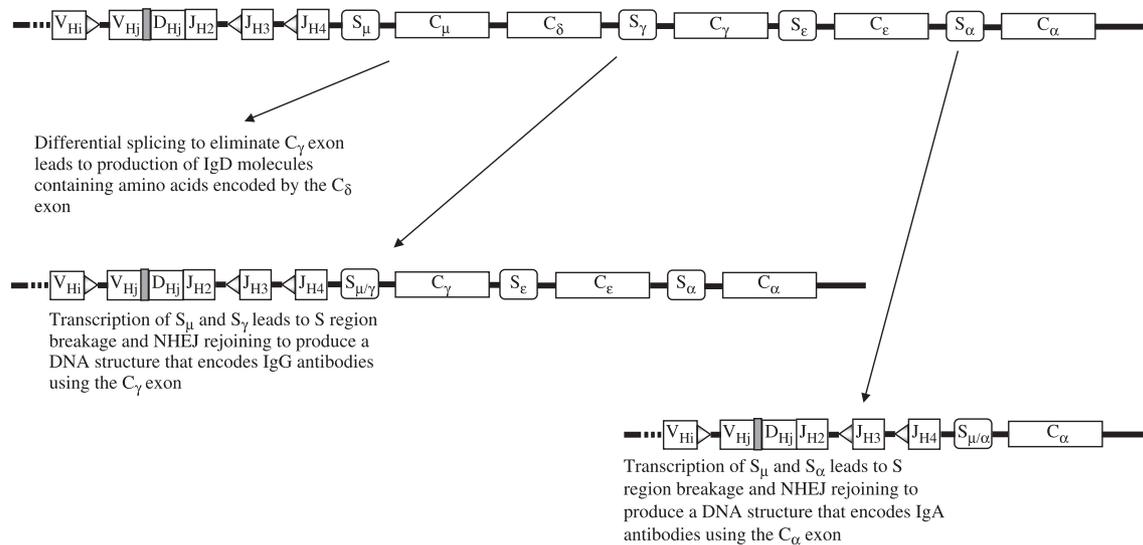


Fig. 4. Differential splicing and class switch rearrangements (CSRs) leading to synthesis of different Ig classes. The details are explained in Kinoshita and Honjo, 2001. To synthesize other immunoglobulin classes instead of IgM molecules, B cells either change Ig mRNA splicing, to incorporate the C<sub>δ</sub> exon for IgD production, or engage in class switch rearrangement (CSR) so that a C<sub>γ</sub>, C<sub>ε</sub>, or C<sub>α</sub> exon for IgG, IgE or IgA production is juxtaposed downstream of the V region sequences. CSR occurs when the switch regions (large rounded rectangles) upstream of C<sub>μ</sub> and another C region exons are transcribed, cleaved and rejoined to make a hybrid S<sub>μ/x</sub> region. Each S region comprises a lymphokine-regulated promoter upstream of a sequence containing many direct and inverted DNA repeats. The actual number of C region exons has been reduced to simplify the figure.

Table 1  
Specificity of natural genetic engineering functions

Example	Observed specificity (mechanism)	References
Mating type cassette switching ( <i>S. cerevisiae</i> )	Localized, directional gene conversion (HO endonuclease cleavage initiates homology-dependent recombination)	Haber, 1998
Immune system V(D)J joining	Cleavage at specific recombination signal sequences (RSSs); flexible joining by non-homologous end joining (NHEJ) functions (recognition of RSSs by RAG1+2 transposase)	Bassing et al., 2002; Gellert, 2002
Immune system somatic hypermutation	5' exons of immunoglobulin determinants (transcriptional specificity)	Kinoshita and Honjo, 2001
Immune system class switching	Lymphokine-controlled choice of switch regions (promoter activation)	Kinoshita and Honjo, 2001
Budding yeast ( <i>S. cerevisiae</i> ) retroviral-like elements Ty1–Ty4	Strong preference for insertion upstream of RNA polymerase III initiation sites (protein–protein interaction of integrase with RNA polymerase III factors)	Kirchner et al., 1995; Kim et al., 1998
Budding yeast retroviral-like element Ty1	Preference for insertion upstream of RNA polymerase II initiation sites rather than exons	Eibel and Philippsen, 1984
Budding yeast retroviral-like element Ty5	Strong preference for insertion in transcriptionally silenced regions of the yeast genome (protein–protein interaction of integrase with Sir4 silencing protein)	Zou et al., 1996; Sandmeyer, 2003; Xie et al., 2001
Fission yeast ( <i>S. pombe</i> ) retroviral-like elements Tf1 and Tf2	Insertion almost exclusively in intergenic regions (>98% for Tf1); biased towards PolIII promoter-proximal sites, 100–400 bp upstream of the translation start; preference for chromosome 3	Singleton and Levin, 2002; Bowen et al., 2003; Kordis, this issue
Murine Leukemia Virus (MLV)	Preference for insertion upstream of transcription start sites in human genome	Wu et al., 2003; Mitchell et al., 2004
HIV	Preference for insertion into actively transcribed regions of human genome	Mitchell et al., 2004
Drosophila P-factors	Preference for insertion into the 5' end of transcripts	Spradling et al., 1995
Drosophila P-factors	Targeting (“homing”) to regions of transcription factor function by incorporation of cognate binding site; region-specific	Hama et al., 1990; Kassir et al., 1992; Fauvarque and Dura, 1993; Taillebourg and Dura, 1999
HeT-A and TART retrotransposons	Insertion at Drosophila telomeres	Pardue and DeBaryshe, 2003
R1 and R2 LINE element retrotransposons	Insertion in arthropod ribosomal 28S coding sequences (sequence-specific endonuclease, reverse transcription)	Xiong and Eickbush, 1988; Burke et al., 1999
Group I homing introns (DNA based)	Site-specific insertion into coding sequences in bacteria and eukaryotes (sequence-specific endonuclease)	Belfort and Perlman, 1995
Group II homing introns (RNA based)	Site-specific insertion into coding sequences in bacteria and eukaryotes (RNA recognition of DNA sequence motifs, reverse transcription)	Mohr et al., 2000; Karberg et al., 2001

The immune system provides one case where cells display how much control they can exert over DNA restructuring. Further examples are found in other instances of developmental DNA rearrangements (e.g. [Beermann, 1977](#); [Wyngaard and Gregory, 2001](#); [Prescott, 2000](#)). B cells further illustrate the potential cells have to turn natural genetic engineering activities on and off in response to internal and external signals (rearranged DNA, immunoglobulin chains, mitogenic antigen binding, lymphokines). They show how DNA rearrangements can be both highly specific, directed by DNA sequences or transcriptional activity, and yet flexible, using untemplated nucleotides and variable internucleotide linkages to enhance combinatorial diversity. The mixture of specificity and flexibility enables the V(D)J joining system to produce extraordinary protein diversity (on the order of  $10^{12}$  combinations) while conserving H and L chain structures. By combining specificity and flexibility, immune system engineering optimizes the chance to produce a functional antibody molecule with an indeterminate specificity. In addition, the synergy of transcriptional regulation and DNA signals seen in CSR provides a comprehensible mechanism for cellular direction of DNA restructuring activities.

It is highly significant that the degree of cellular control over natural genetic engineering exemplified by lymphocytes and other developmental systems is not an isolated case. Experimentation with a number of different mobile element systems has shown that they can be activated temporarily by response to particular conditions. The conditions are quite varied, ranging from blockage of normal chromosome separation during early embryonic development ([McClintock, 1987](#)) to osmotic and other physical stresses associated with protoplast regeneration ([Wessler, 1996](#)) to oxidative starvation stress during “adaptive mutation” ([Shapiro, 1984](#); [Hall, 1988](#); [Maenhaut-Michel and Shapiro, 1994](#); [McKenzie et al., 2000](#); [Ilves et al., 2001](#)) to mating outside the normal breeding group causing “hybrid dysgenesis” ([Bregliano and Kidwell, 1983](#); [Engels, 1989](#); [O’Neill et al., 1998](#); [Vrana et al., 2000](#)). In the case of hybrid dysgenesis, it is important that changes induced typically occur during the mitotic development of the germ line ([Woodruff and Thompson, 2002](#)). Since the progeny of germ line cells that have undergone DNA rearrangements produce multiple gametes, mating-induced natural genetic engineering can lead to the appearance of small interbreeding populations carrying similarly restructured genomes.

In addition to control over when and in what situations natural genetic engineering functions become active, there are a variety of examples where mobile elements display various degrees of targeting specificity in the genome (Table 1). In some cases, we know the molecular basis for targeting. Connections between DNA rearrangement specificity, on the one hand, and transcriptional control or chromatin formatting functions, on the other, are particularly noteworthy for the following reason. Most biologists recognize that signal transduction networks can direct transcriptional and chromatin formatting activities to particular regions or sites in the genome. Thus, connecting these activities to the operation of mobile elements establishes a readily understood mechanistic basis for cellular control networks targeting DNA rearrangements in response to internal and external signals.

## 5. Conclusions: a 21st century view of evolution

Based on discoveries about genome system architecture and natural genetic engineering, it is now possible to formulate a series of basic concepts that lead to viewing evolution as something akin to a systems engineering process:

- 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 sophisticated cellular natural genetic engineering systems operating non-randomly.
- Significant evolutionary changes can result from altering the repetitive elements formatting genome system architecture, not just from altering protein and RNA coding sequences.
- Cellular regulation of natural genetic engineering activities makes evolutionary change responsive to biological inputs with respect to timing and location of DNA rearrangements.

These basic ideas about the role of cell-regulated natural genetic engineering of genome system architecture have implications for how we think about the evolutionary process, and previous articles have discussed some of these (Shapiro, 1999, 2002a; Shapiro and Sternberg, 2005). In the context of this symposium, it is worthwhile to emphasize how natural genetic engineering (i) can increase the efficiency of searching for genome configurations that encode functional complex systems and (ii) can favor the elaboration of hierarchic system architectures.

As we saw in the immune system example, natural genetic engineering takes existing functional coding modules and assembles them into new combinations. Since the rearranged DNA segments already have functionality, the potential of the newly assembled genomic structure for

adaptive utility is greater than for a structure resulting from random changes. The same is true of other examples of natural genetic engineering. For example, insertion of a mobile element containing a package of integrated transcription and chromatin-formatting signals can place an existing coding region under novel controls. In this way, a working product can be expressed under conditions where it was previously absent (e.g. Errede et al., 1981). The evidence is quite solid that this process has taken place during evolution (Britten, 1997; Brosius, 1999; Jordan et al., 2003), and transcript profiling during mouse oocyte development indicates that retroviral promoters regulate expression of many embryonic functions (Peaston et al., 2004). Similarly, insertion of a DNA segment encoding a functional domain is more likely to add new capabilities to a protein than are random changes in sequence or addition of random polypeptide components. Domain addition is commonly used in laboratory engineering of proteins.

Acquisition of new DNA regulatory regions and protein domains are examples of engineering a new system by arranging known components in new combinations. The rearrangement process can always be followed, as it often is in human engineering, by fine-tuning or modification of individual components (microevolution). Here again, the immune system is instructive. A similar “rearrangement-followed-by-fine-tuning” sequence of events occurs in targeted somatic hypermutation of joined exons encoding antigen-binding domains of immunoglobulins (Bassing et al., 2002; Kinoshita and Honjo, 2001).

The ability to regulate DNA rearrangements in time and location within the genome also adds significantly to the evolutionary efficiency of genome restructuring. By making sure that genomes in normally reproducing organisms are stable and that the genomes of cells under stress are mutable, networks activating natural genetic engineering functions provide hereditary variability when it is most needed (McClintock, 1984). Episodic activation of genome restructuring functions means that multiple changes can occur when complex rearrangements may be required to meet adaptive needs. It further predicts that evolutionary change will be inherently intermittent and punctuated rather than continuous (cf. Gould and Eldredge, 1993).

Targeting of genetic change has potential advantages. Restricting somatic hypermutation in B cells to exons for antigen binding domains and targeting of retrotransposon insertions to the upstream regions of genetic loci (Table 1) are obvious examples. Restricting retrotransposon insertion sites minimizes interruption of coding sequences and thus, presumably, enhances the potential for a constructive regulatory change. In yeast, selections for increased protein expression most commonly produce mutant strains carrying just such retrotransposon insertions (Errede et al., 1981). Although virtually every mobile element displays some degree of target selectivity, there have been no careful studies of how selectivity may influence the ability of the element to make useful changes. Now that experimenters

are learning how to target mobile elements (Table 1; Mohr et al., 2000; Karberg et al., 2001; Sandmeyer, 2003; Zhu et al., 2003), the time is ripe to investigate whether enhanced targeting alters their ability to generate adaptive changes.

In addition to increasing the efficiency of genome restructuring in response to challenge, the action of natural genetic engineering systems also imparts structural characteristics to genomes. Duplication and rearrangement of genomic segments can involve DNA sequences in the megabase range (Harden and Ashburner, 1990; Bailey et al., 2003). So natural genetic engineering has the potential to facilitate the establishment and amplification of higher order genomic subsystems, as has clearly occurred in the evolution of homeodomain complexes (Patel and Prince, 2000). This tendency to amplify progressively larger subsystems may help explain the hierarchic nature of genome coding (Van Driel et al., 2003).

Another result of change by natural genetic engineering is the tendency for genomes to accumulate dispersed copies of repeats. This tendency is usually explained by the “selfish DNA” hypothesis (Doolittle and Sapienza, 1980; Orgel and Crick, 1980). However, the selfish DNA view does not take into consideration the well-documented functional information found in all classes of repetitive DNA elements (Shapiro and Sternberg, 2005). Since dispersed repeats influence both coding sequence expression and physical organization of genomes, an alternative functionalist hypothesis must be entertained: namely, that repeat distribution reflects the establishment of a system architecture required for effectively integrated genome functioning. While it is difficult right now to test these alternative models experimentally, formal investigation remains possible. Employing simulated evolutionary processes, such as genetic programming (Koza and Andre, 1996), computer scientists can test whether the presence of mobile formatting repeats speeds up the evolutionary process. A positive result in the simulation would be a strong spur to the development of experimental systems.

It appears from this discussion that a distinct 21st Century view of evolution can stimulate research at the interface between experimental observation-based biology and mathematical analysis of complex systems. That was the objective of the present symposium. The ideas presented here are consistent with molecular genetics but are quite different from conventional evolutionary theory. Whether they prove to be predictive or not remains unknown until tests have been performed. Nonetheless, the value of opening up scientific exploration of evolution to new fundamental concepts should be clear. It will lead us to ask questions that could not have been imagined in the middle of the 20th century.

## References

Agrawal, A., Eastman, Q.M., Schatz, D.G., 1998. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* 394, 744–751.

- Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815.
- Bailey, J.A., Liu, G., Eichler, E.E., 2003. An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* 73, 823–834.
- Bassing, C.H., Swat, W., Alt, F.W., 2002. The mechanism and regulation of chromosomal V(D)J recombination. *Cell* 109, S45–S55.
- Beermann, S., 1977. The diminution of heterochromatic chromosomal segments in *Cyclops* (Crustacea, Copepoda). *Chromosoma* 60, 297–344.
- Belfort, M., Perlman, P.S., 1995. Mechanism of intron mobility. *J. Biol. Chem.* 270, 30237–30240.
- Bennett, P., 2000. Demystified . . . microsatellites. *MP, Mol. Pathol.* 53, 177–183.
- Berg, D.E., Howe, M.M. (Eds.), *Mobile DNA*. ASM Press, Washington, DC.
- Blackburn, E.H., 2001. Switching and signaling at the telomere. *Cell* 106, 661–673.
- Bowen, N.J., Jordan, I.K., Epstein, J.A., Wood, V., Levin, H.L., 2003. Retrotransposons and their recognition of pol II promoters: a comprehensive survey of the transposable elements from the complete genome sequence of *Schizosaccharomyces pombe*. *Genome Res.* 13, 1984–1997.
- Bregliano, J.C., Kidwell, M., 1983. Hybrid dysgenesis. In: Shapiro, J.A. (Ed.), *Mobile Genetic Elements*. Academic Press, New York, pp. 363–410.
- Britten, R.J., 1997. Mobile elements inserted in the distant past have taken on important functions. *Gene* 205, 177–182.
- Brosius, J., 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* 238, 115–134.
- Bukhari, A.I., Shapiro, J.A., Adhya, S.L., 1977. *DNA Insertion Elements, Episomes and Plasmids*. Cold Spring Harbor Press, Cold Spring Harbor, NY.
- Burke, W.D., Malik, H.S., Jones, J.P., Eickbush, T.H., 1999. The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. *Mol. Biol. Evol.* 16, 502–511.
- Choo, K.H., 2001. Domain organization at the centromere and neocentromere. *Dev. Cell* 1, 165–177.
- Coffin, J.M., Hughes, S.H., Varmus, H.E., 1997. *Retroviruses*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Craig, N.L., Craigie, R., Gellert, M., Lambowitz, A.M., 2002. *Mobile DNA II*. ASM Press, Washington, DC.
- Csirik, A.K., Henikoff, S., 1998. Something from nothing: the evolution and utility of satellite repeats. *Trends Genet.* 14, 200–204.
- Deininger, P.L., Moran, J.V., Batzer, M.A., Kazazian, H.H., 2003. Mobile elements and mammalian genome evolution. *Curr. Opin. Genet. Dev.* 13, 651–658.
- Donehower, L., Gillespie, D., 1979. Restriction site periodicities in highly repetitive DNA of primates. *J. Mol. Biol.* 134, 805–834.
- Doolittle, R.F., 1995. The multiplicity of domains in proteins. *Ann. Rev. Biochem.* 64, 287–314.
- Doolittle, W.F., Sapienza, C., 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284, 601–603.
- Dowsett, A.P., 1983. Closely related species of *Drosophila* can contain different libraries of middle repetitive DNA sequences. *Chromosoma* 88, 104–108.
- Eibel, H., Philippsen, P., 1984. Preferential integration of yeast transposable element Ty into a promoter region. *Nature* 307, 386–388.
- Eichler, E.E., 2001. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* 17, 661–669.
- Engels, W.R., 1989. P Elements in *Drosophila Melanogaster*. In: Berg, D.E., Howe, M.M. (Eds.), *Mobile DNA*. ASM Press, Washington, DC, pp. 437–484. <http://engels.genetics.wisc.edu/Pelements/index.html>.
- Errede, B., Cardillo, T.S., Wever, G., Sherman, F., 1981. ROAM mutations causing increased expression of yeast genes: their activation by signals directed toward conjugation functions and their formation by insertions

- of Ty1 repetitive elements. Cold Spring Harbor Symp. Quant. Biol. 45, 593–607.
- Fauvarque, M.O., Dura, J.M., 1993. Polyhomeotic regulatory sequences induce developmental regulator-dependent variation and targeted P-element insertions in *Drosophila*. *Genes Dev.* 7, 1508–1520.
- Gellert, M., 2002. V(D)J recombination: RAG proteins, repair factors, and regulation. *Ann. Rev. Biochem.* 71, 101–132.
- Gould, S.J., Eldredge, N., 1993. Punctuated equilibrium comes of age. *Nature* 366, 223–227.
- Haber, J.E., 1998. Mating-type gene switching in *Saccharomyces cerevisiae*. *Annu. Rev. Genet.* 32, 561–599.
- Hall, B.G., 1988. Adaptive evolution that requires multiple spontaneous mutations: I. Mutations involving an insertion sequence. *Genetics* 120, 887–897.
- Hama, C., Ali, Z., Kornberg, T.B., 1990. Region-specific recombination and expression are directed by portions of the *Drosophila* engrailed promoter. *Genes Dev.* 4, 1079–1093.
- Harden, N., Ashburner, M., 1990. Characterization of the FB-NOF transposable element of *Drosophila melanogaster*. *Genetics* 126, 387–400.
- Iives, H., Horak, R., Kivisaar, M., 2001. Involvement of sigma(S) in starvation-induced transposition of *Pseudomonas putida* transposon Tn4652. *J. Bacteriol.* 183, 5445–5448.
- International Human Genome Consortium, 2001. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
- Jacob, F., Monod, J., 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3, 318–356.
- Jacob, F., Brenner, S., Cuzin, F., 1963. On the regulation of DNA replication in Bacteria. Cold Spring Harbor Symp. Quant. Biol. 28, 329–438.
- 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 Genet.* 19, 68–72.
- Karberg, M., Guo, H., Zhong, J., Coon, R., Perutka, J., Lambowitz, A.M., 2001. Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. *Nat. Biotechnol.* 19, 1162–1167.
- Kassis, J.A., Noll, E., Vansickle, E.P., Odenwald, W.F., Perrimon, N., 1992. Altering the insertional specificity of a *Drosophila* transposable element. *Proc. Natl. Acad. Sci. U. S. A.* 89, 1919–1923.
- Kazazian Jr., H.H., 2000. Genetics. L1 retrotransposons shape the mammalian genome. *Science* 289, 1152–1153.
- Kim, J.M., Vanguri, S., Boeke, J.D., Gabriel, A., Voytas, D.F., 1998. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res.* 8, 464–478.
- Kinoshita, K., Honjo, T., 2001. Linking class-switch recombination with somatic hypermutation. *Nat. Rev., Mol. Cell Biol.* 2, 493–503.
- Kirchner, J., Connolly, C.M., Sandmeyer, S.B., 1995. Requirement of RNA polymerase III transcription factors for in vitro position-specific integration of a retroviruslike element. *Science* 267, 1488–1491.
- Koza, J.R., Andre, D., 1996. A case study where biology inspired a solution to a computer science problem. *Pac. Symp. Biocomput.*, 500–511.
- Lerman, D.N., Michalak, P., Helin, A.B., Bettencourt, B.R., Feder, M.E., 2003. Modification of heat-shock gene expression in *Drosophila melanogaster* populations via transposable elements. *Mol. Biol. Evol.* 20, 135–144.
- Maenhaut-Michel, G., Shapiro, J.A., 1994. The roles of starvation and selective substrates in the emergence of *araB-lacZ* fusion clones. *EMBO J.* 13, 5229–5239.
- 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.
- McKenzie, G.J., Harris, R.S., Lee, P.L., Rosenberg, S.M., 2000. The SOS response regulates adaptive mutation. *Proc. Natl. Acad. Sci. U. S. A.* 97, 6646–6651.
- Mitchell, R.S., Beitzel, B.F., Schroder, A.R.W., Shinn, P., Chen, H., et al., 2004. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.* 2, e234.
- Mohr, G., Smith, D., Belfort, M., Lambowitz, A.M., 2000. Rules for DNA target-site recognition by a lactococcal group II intron enable retargeting of the intron to specific DNA sequences. *Genes Dev.* 14, 559–573.
- Moran, J.V., DeBerardinis, R.J., Kazazian Jr., H.H., 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.
- Nekrutenko, A., Li, W.-H., 2001. Transposable elements are found in a large number of human protein coding regions. *Trends Genet.* 17, 619–625.
- O'Neill, R.J., O'Neill, M.J., Graves, J.A., 1998. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* 393, 68–72.
- Orgel, L.E., Crick, F.H., 1980. Selfish DNA: the ultimate parasite. *Nature* 284, 604–607.
- Pardue, M.L., DeBaryshe, P.G., 2003. Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. *Annu. Rev. Genet.* 37, 485–511.
- Patel, N.H., Prince, V.E., 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, A.E., et al., 2004. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev. Cell* 7, 597–606.
- Prescott, D.M., 2000. Genome gymnastics: unique modes of DNA evolution and processing in Ciliates. *Nat. Rev., Genet.* 1, 191–198.
- Reznikoff, W.S., 1992. The lactose operon-controlling elements: a complex paradigm. *Mol. Microbiol.* 6, 2419–2422.
- Sandmeyer, S.B., 2003. Integration by design. *Proc. Natl. Acad. Sci. U. S. A.* 100, 5586–5588.
- Schotta, G., Ebert, A., Dorn, R., Reuter, G., 2003. Position-effect variegation and the genetic dissection of chromatin regulation in *Drosophila*. *Semin. Cell Dev. Biol.* 14, 67–75.
- Shapiro, J., 1979. A molecular model for the transposition and replication of bacteriophage Mu and other transposable elements. *Proc. Natl. Acad. Sci. U. S. A.* 76, 1933–1937.
- Shapiro, J.A., 1983. Mobile Genetic Elements. Academic Press, New York, NY.
- Shapiro, J.A., 1984. Observations on the formation of clones containing *araB-lacZ* cistron fusions. *MGG, Mol. Gen. Genet.* 194, 79–90.
- Shapiro, J.A., 1992. Natural genetic engineering in evolution. *Genetica* 86, 99–111.
- Shapiro, J.A., 1997. Genome organization, natural genetic engineering, and adaptive mutation. *Trends Genet.* 13, 98–104.
- Shapiro, J.A., 1999. Genome system architecture and natural genetic engineering in evolution. In: Caporale, L. (Ed.), *Molecular Strategies for Biological Evolution*, Ann. N.Y. Acad. Sci., vol. 870, pp. 23–35.
- Shapiro, J.A., 2002a. Genome organization and reorganization in evolution: formatting for computation and function. In: Speybroeck, L.V., de Vijver, G.V., de Waele, D. (Eds.), *From Epigenesis to Epigenetics: The Genome in Context*, Ann. N.Y. Acad. Sci., vol. 981, pp. 111–134.
- Shapiro, J.A., 2002b. A 21st Century view of evolution. *J. Biol. Phys.* 28, 745–764.
- Shapiro, J.A., Sternberg, R.V., 2005. Why repetitive DNA is essential for genome function. *Biol. Rev.* (in press).
- Singleton, T.L., Levin, H.L., 2002. A long terminal repeat retrotransposon of fission yeast has strong preferences for specific sites of insertion. *Eukaryot. Cell* 1, 44–55.
- Spofford, J.B., 1976. Position-effect variegation in *Drosophila*. In: Ashburner, M., Novitski, E. (Eds.), *The Genetics and Biology of Drosophila*. Academic Press, New York, NY, pp. 955–1018.
- Spradling, A.C., Stern, D., Kiss, I., Roote, J., Lavery, T., Rubin, G.M., 1995. Gene disruptions using P transposable elements. *Proc. Natl. Acad. Sci. U. S. A.* 92, 10824–10830.

- Sternberg, R.V., Shapiro, J.A., 2005. How repeated retroelements format genome function. *Cytogenet. Genome Res.* (in press).
- Taillebourg, E., Dura, J.M., 1999. A novel mechanism for P element homing in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* 96, 6856–6861.
- Turing, A.M., 1950. Computing machinery and intelligence. *Mind (J. Mind Assoc.)* 59, 433–460.
- van Driel, R., Fransz, P.F., Verschure, P.J., 2003. The eukaryotic genome: a system regulated at different hierarchical levels. *J. Cell. Sci.* 116, 4067–4075.
- Vrana, P.B., Fossella, J.A., Matteson, P.G., O'Neill, M.J., Tilghman, S.M., 2000. Genetic and epigenetic incompatibilities underlie hybrid dysgenesis in *peromyscus*. *Nat. Genet.* 25, 120–124.
- Wessler, S.R., 1996. Turned on by stress: plant retrotransposons. *Curr. Biol.* 6, 959–961.
- Woodruff, R.C., Thompson Jr., J.N., 2002. Mutation and premating isolation. *Genetica* 116, 371–382.
- Wu, X., Li, Y., Crise, B., Burgess, S.M., 2003. Transcription start regions in the human genome are favored targets for MLV integration. *Science* 300, 1749–1751.
- Wyngaard, G.A., Gregory, T.R., 2001. Temporal control of DNA replication and the adaptive value of chromatin diminution in copepods. *J. Exp. Zool.* 291, 310–316.
- Xie, W., Gai, X., Zhu, Y., Zappulla, D.C., Sternglanz, R., Voytas, D.F., 2001. Targeting of the yeast Ty5 retrotransposon to silent chromatin is mediated by interactions between integrase and Sir4p. *Mol. Cell. Biol.* 21, 6606–6614.
- Xiong, Y., Eickbush, T.H., 1988. Functional expression of a sequence-specific endonuclease encoded by the retrotransposon R2Bm. *Cell* 55, 235–246.
- Zadeh, L., 1975. The calculus of fuzzy restrictions. In: Zadeh, L.A., et al., (Eds.), *Fuzzy Sets and Applications to Cognitive and Decision Making Processes*. Academic Press, New York, NY, pp. 1–39.
- Zdobnov, E.M., et al., 2002. Comparative genome and proteome analysis of *anopheles gambiae* and *drosophila melanogaster*. *Science* 298, 149–159.
- Zhu, Y., Dai, J., Fuerst, P.G., Voytas, D.F., 2003. Controlling integration specificity of a yeast retrotransposon. *Proc. Natl. Acad. Sci. U. S. A.* 100, 5891–5895.
- Zou, S., Ke, N., Kim, J.M., Voytas, D.F., 1996. The *Saccharomyces retrotransposon* Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci. *Genes Dev.* 10, 634–645.