Genome Informatics: The Role of DNA in Cellular Computations

James A. Shapiro
Department of Biochemistry and Molecular Biology
University of Chicago, IL, USA
jsha@uchicago.edu

Abstract

Cells are cognitive entities possessing great computational power. DNA serves as a multivalent information storage medium for these computations at various time scales. Information is stored in sequences, epigenetic modifications, and rapidly changing nucleoprotein complexes. Because DNA must operate through complexes formed with other molecules in the cell, genome functions are inherently interactive and involve two-way communication with various cellular compartments. Both coding sequences (data files) and repetitive sequences (generic formatting signals) contribute to the hierarchical systemic organization of the genome. By virtue of nucleoprotein complexes, epigenetic modifications, and natural genetic engineering activities, the genome can serve as a read–write storage system. An interactive informatic conceptualization of the genome allows us to understand the functional importance of DNA that does not code for protein or RNA structure, clarifies the essential multidirectional and systemic nature of genomic information transfer, and emphasizes the need to investigate how cellular computation operates in reproduction and evolution.

Keywords

computation, evolution, formatting, genome system architecture, natural genetic engineering, repetitive DNA

September 4, 2006; accepted October 11, 2006

The early pioneers of molecular biology believed that the new science would provide a solid physical and chemical basis for the mechanistic views of heredity and cell function that prevailed in mid-20th century (Judson 1979). It is, therefore, a great irony that molecular analysis has led biology into the informatic realm of complexity, redundancy, signaling, networks, and decision making (Gerhart and Kirschner 1997; Alberts et al. 2002). As inevitably happens in science, new technology uncovers new phenomena that require new concepts (Kuhn 1962). The old atomistic pre-DNA concepts of genome action and genotype-phenotype relationships are no longer capable of explaining or synthesizing the tsunami of data we now confront in cell, developmental, and evolutionary biology. This review is an attempt to bring together some fundamental insights from the past six decades and to suggest novel conceptual bases for describing genome function that may help clarify a bewildering situation.

1. The Meaning of Genome Informatics in a Cognitive Context

A basic lesson from five decades of molecular biology is that cells are immensely sophisticated cognitive and computational entities. This realization comes from many sources: studies of cellular metabolism and its control, analysis of basic processes like DNA replication and transcription, molecular dissection of cell cycle regulation, investigation of cellular differentiation during multicellular development, elucidation of cellular mechanisms for damage detection and repair, studies of cell motility, and so on. It is difficult to find a basic cell process that does not involve one or more sensory event(s) followed by processing of the sensory input(s). During DNA replication, to take a very simple example, error correction begins when a proofreading exonuclease or a MutS-type protein senses helix distortion due to base mismatching (Kunkel and Erie 2005). When we shift our thinking from individual biochemical processes to the informatics of cell proliferation, the magnitude of cellular processing capabilities becomes even more apparent. A tremendous cybernetic challenge arises during every cell cycle from the need to keep millions of biochemical and biomechanical operations under control in changing conditions. Cellular monitoring and regulatory systems continuously receive multiple inputs containing information about factors such as the status of genome replication, where the cell is in the cell cycle, what nutrients are available, the integrity of supramolecular structures, what intercellular signaling molecules are present, and what other cells are touching the cell surface. With remarkable reliability, the complex information in these inputs is evaluated and processed so that the appropriate molecular events ensue to facilitate cell survival, cell proliferation, cell differentiation, or (when needed) cell death. It is difficult to overstate the fundamental importance of sensory inputs and information processing in maintaining living systems. The tremendous expenditure of high-energy phosphodiester bonds in production and turnover of RNA molecules and in protein modification cycles indicates that information is probably of far higher value in cellular economics than free energy stored in chemical structures like ATP. General considerations such as these dictate the need for more informatics-based concepts of genome action.

The term "computation" is used here to denote information processing that produces functional outputs. Computational concepts such as intercellular signaling, intracellular signal transduction, and checkpoints have been found to apply to all realms of cell action (Hartwell 1992; Gerhart and Kirschner 1997; Alberts et al. 2002). The closest published parallels to the concept of cellular computation advanced here come from neurobiologists who recognize computational ability in individual neurons (London and Hausser 2005; Sidiropoulou et al. 2006). The computer metaphor serves to place the analysis of how cells evaluate multiple inputs and decide appropriate outputs in a scientific context, something that was virtually impossible before the development of electronic information-processing systems. Nonetheless, it is important to keep in mind that cellular computation operates by different principles from most electronic computers. In particular, cellular computation is largely a parallel and distributed analogue process (Bray 1990), not a sequential linear and digital process as classically defined (Turing 1950). Analogies to electronic hardware and software are used without implying any simple correspondence between biological and electronic processes, and a number of fundamental differences between the two modes of information processing will be emphasized.

The phrase "genome informatics" refers to the various roles that DNA molecules play in cellular computations. This connotation differs fundamentally from the more common use of the phrase to mean computer analysis of DNA sequence information. In this article, the emphasis is placed on genome function in the context of the cell as a complex information-processing entity. The purpose is to review some basic principles of genome informatics that may have been undervalued and that may lead to fresh ways of thinking about genome organization and its reorganization in evolution. This informatic perspective offers three particular advantages over classical genetic concepts: (i) it provides a fundamental and an essential role for so-called noncoding DNA sequences; (ii) it makes multidirectional information transfer and systemic integration obligatory for genome functioning; and (iii) it elucidates how cellular computations not only guide genome expression and transmission but may also influence genome evolution.

2. What Roles Does DNA Play in Cellular Informatics?

Rather than a master blueprint for phenotype, a more appropriate metaphor is to think of DNA as an information storage medium accessible by cellular computing networks. Some aspects of this comparison were considered by Atlan and Koppel (1990). The data stored in the genome are necessary to execute cellular tasks. But genomic information alone is not sufficient for those tasks. It must be accessed in the appropriate cellular context to be utilized successfully. This contingent historical aspect of genome function is clearly evident, for example, in progress through the many coordinated events of the cell cycle (Alberts et al. 2002; Murray 2004). Some tasks involving the genome can be executed by accessing different alternative combinations of stored data. The distributed nature of network function underpins the robustness of many characters to mutational damage and explains why so many knockouts have no obvious mutant phenotype (Bray 1990).

2.1. Cellular and Organismal Phenotypes Are Not Hard-wired in Genomes

A major task in reconceptualizing genome function is to divest ourselves of genetic reductionism: the idea of DNA dictating phenotypes and "determining" particular traits of individual genes. While widespread, this view is incompatible with longstanding facts showing that genome function is highly context dependent and that the same DNA molecules function to support very diverse cell and organismal types. In organisms with complex life cycles, dramatically different life forms share the same genome, such as a caterpillar and the butterfly it will morph into. In multicellular organisms (and even in many bacteria; Shapiro and Dworkin 1997), different cell types form without any change in genome content. Genome conservation in differentiated cells is what makes cloning by nuclear transplantation possible (Galli et al. 2003; Gurdon et al. 2003; Wilmut and Paterson 2003). The phenomena of regulation and induction in response to experimental manipulation of animal development (Driesch 1908; Spemann and Mangold 1924; Spemann 1938) indicate that many steps in multicellular ontogeny are often dramatically dependent on nongenomic inputs. The tendency toward genetic reductionism is particularly marked in fields where there has been major progress on clarifying genomic contributions, such as embryogenesis, but even in these fields we know that the roles played by other cellular and organismal components, such as maternal proteins and RNAs in fertilized eggs (Nusslein-Volhard 1991; Gao and Latham 2004), have yet to be fully investigated.

2.2. DNA Is a Multivalent and Interactive Information Storage Medium

The information storage role of DNA is considerably more complex than imagined in the early days of molecular biology, when only sequence data were thought to be significant (Watson and Crick 1953; Benzer 1962). DNA stores information in at least three different forms, each of which operates at a distinct biological time scale.

2.2.1. Genetic storage Information is stored in nucleotide sequences over many cell and organism generations, including (but not limited to) the data files for the primary structures of protein and RNA products. Other important classes of sequence information include the repetitive signals needed to direct cellular activity on the genome (Shapiro and von Sternberg 2005). Sequence information is the most widely recognized way that genomes hold information, and it is the most stable form of storage. In conventional theory, DNA sequence information is implicitly considered read-only memory (ROM) storage, a hard-wired part of the system that changes only by accidents and malfunctioning of the replication machinery. However, it is better to consider DNA sequence information as equivalent to data stored on a hard disk. Like magnetically stored information, sequence data is subject to modification, and we shall see that cells have the biochemical hardware needed to rewrite DNA sequences. From this perspective, genetic storage can operate as a read-write (RW) memory, as do epigenetic and computational storage.

2.2.2. Epigenetic storage Information can be stored over multiple cell generations in the form of covalent modifications to specific residues, such as cytosine methylation (Bird 2002), and also as heritable "chromatin" complexes involving proteins and RNA (Henikoff and Ahmed 2005; Bernstein and Allis 2005). This metastable form of storage is generally referred to as "epigenetic inheritance" (Jaenisch and Bird 2003). Covalent modification patterns and chromatin configurations can be maintained through many cell cycles, but they are also subject to active and rapid change by cellular "chromatin remodeling" machinery (Muller and Leutz 2001). Specific regions of the genome can be independently remodeled. Multicomponent chromatin complexes compact the DNA in various ways and control its availability to the molecules responsible for replication, transcription, and other processes (Zaidi et al. 2005). When particular genetic loci are remodeled during gamete formation and the effects are observed in their progeny, the loci displaying parent-specific expression patterns are said to be "imprinted" (Mann 2001; Li 2002). Recent evidence indicates that modification of epigenetic storage underlies cellular differentiations during multicellular development (Bibikova et al. 2006). The role of epigenetic changes in forming differentiated adult cells clarifies much of the phenomenology associated with cloning by nuclear transplantation (Galli et al. 2003; Gurdon et al. 2003; Wilmut and Paterson 2003).

2.2.3. Computational storage Information about recent conditions inside and outside the cell is maintained in the form of transient nucleoprotein complexes reflecting recent responses to internal and external signals. These complexes represent the genomic nodes of signal transduction networks and can change rapidly as particular signals increase or decrease in intensity. Thus, there is a short-term, highly dynamic form of information storage (analogous to RAM memory) that reflects the current status of the physical, nutritional, and biological environment and that also represents internal processes, such as cell growth and progress through the cell cycle.

2.3. DNA Is a Substrate for Nucleoprotein Complexes

The most basic fact incompatible with conventional ideas about genome-directed phenotypes is that DNA, by itself, does very little in living cells. The major genomic processes of compaction, replication, transcription, transmission to daughter cells, repair and restructuring all involve complexes between DNA and other cell molecules. It is impossible to overstate the importance of bringing our concepts into alignment with this fundamental molecular reality. Analyzing the formation and turnover of nucleoprotein complexes is central to understanding genome informatics. On this basis alone, models that do not incorporate this interactive process and its potential for information transfer are unrealistic. There are different molecular bases underlying specific nucleoprotein formation.

2.3.1. RNA-guided nucleoprotein complex formation RNA recognition and binding involves double-stranded (DS) small interfering (si-) and micro (mi-) RNAs about two dozen base-pairs long, with sequence complementarity determining specificity on the DNA (Almeida and Allshire 2005; Bernstein and Allis 2005). Elimination of the machinery for producing short DS RNA molecules blocks the production of mi-RNA regulatory molecules and disrupts complexes that direct the formation of silent chromatin regions. Such silenced chromatin controls transcription and plays an essential role in chromosome behavior during the cell cycle (Hall et al. 2003). The requirement for DS RNA as a substrate for the si- and mi-RNA processing machinery directs initiation of chromatin complexes to sites where both strands of the DNA are transcribed, often at particular repeat sequences such as transposable elements (Lipmann et al. 2004).

2.3.2. Protein-guided nucleoprotein complex formation Protein binding to DNA is based upon the recognition of consensus sequence motifs, as first determined for the binding of *lac* and lambda repressors to their respective operators (Ptashne 1986). The sequence motifs recognized by proteins active on DNA molecules (transcription factors, replication initiation and termination factors, site-specific recombinases, endo- and exonucleases) are sometimes unique in a genome (e.g., the HO endonuclease cleavage site in the *S. cerevisiae* MAT locus where mating-type switching initiates; Haber 1998). However, the general rule is that protein binding sites are repeated. Sometimes the binding site repeats are located at a single locus, (e.g., the palindromic iterated *lac* and lambda operators that promote cooperative formation of tightly localized nucleoprotein complexes; Ptashne 1986; Shapiro and von Sternberg 2005). More frequently, repetitive binding sites are widely distributed in the genome (e.g., sequences stimulating homologous recombination; Smith 1994). Protein-determined specificity in nucleoprotein complex formation arises when distributed binding sites for one protein are combined with binding sites for other proteins to generate organized substrates for the cooperative formation of complexes involving several different DNA-binding factors (e.g., developmental regulatory enhancer regions; Arnone and Davidson 1997; Davidson 2001).

2.3.3. Dynamics of nucleoprotein complexes Complexes formed on a DNA segment comprising multiple binding sites can change their structures as the concentrations of the different protein factors go up and down or as individual components are degraded or chemically modified (e.g., DNA methylation; RNA cleavage; protein phosphorylation, methylation, acetylation). The resulting plasticity in the nucleoprotein complex structure endows these manifold combinatorial binding regions with the ability to participate in nonlinear responses to changing conditions during complex biological processes, such as progress through the cell cycle, adaptation to a changing environment, response to cell or organismal damage, and cellular differentiation during multicellular development (e.g., Yuh et al. 1998; Davidson 2001).

2.4. DNA Participates Actively in Nucleoprotein Complex Formation and Function

It is widely accepted that the dynamics of nucleoprotein complex formation and breakdown provide a key mechanistic basis for cellular computations involving the genome (Alberts et al. 2002). In applying informatic metaphors to these processes, it has been common to employ the Turing distinction between machine and tape (Turing 1950). This distinction assumes that DNA serves only as a carrier of information digitized in nucleotide sequences. Nonetheless, several considerations tell us that DNA is more than a passive coded tape in genomic computations and that it plays an active functional role (Box 1).

Box 1. Direct involvement of DNA in cellular functions: *The placement of binding motifs along the DNA affects the structure and activity of nucleoprotein complexes.* The importance of the order and spacing of protein binding sites was first elucidated in studies of site-specific recombination (Landy 1989) and transcriptional regulation (Ptashne 1986). This means that the DNA is an essential structural component of the active complex, not just a digital coding medium.

The structural role of DNA in nucleoprotein complex formation endows it with allosteric properties and therefore with the ability to operate as a communication molecule. Examples of DNA allostery include cases where binding one protein influences the binding of a second protein because of a change in DNA bending (Cases and de Lorenzo 1998), local changes in superhelical density alter protein binding (Wang and Syvanen 1992; Mukelishvili and Travers 2003), and chromatin domains form sequentially along a chromosome (Razin et al. 2004).

In certain genomic transactions, the DNA plays a direct biochemical role. Examples include the priming of repair synthesis and reverse transcription, strand invasion during homologous recombination, and nucleophilic attacks by 3′-hydroxyl groups in transposition, VDJ recombination and other DNA rearrangements (Craig et al. 2002).

DNA appears to partner Watson-Crick base-pairing in processes directed by si- or mi-RNAs (Bernstein and Allis 2005). The formation of short R-loops in the DNA will distort the helical structure and influence how proteins bind to adjacent DNA, which may be part of the mechanism underlying si- and mi-RNA directed regulatory effects.

Certain exceptional DNA configurations play an important role in spatially organizing the genome within the nucleus or nucleoid. Examples include G quartet structures at telomeres (Williamson 1994) and Z DNA segments in particular repetitive regions (Rothenburg et al. 2001). Other less well studied configurations, like hemicatenanes (Stros et al. 2004), are likely to play additional roles in genome organization. Hemicatenanes directly connect different DNA molecules and can facilitate important processes, such as synapsis of homologous duplex regions during the cell cycle. Such a facilitating effect upon recombinational repair of double-strand breaks may be why hemicatenanes form upon the arrest of DNA synthesis (Lucas and Hyrien 2000).

2.5. DNA Formatting, Hierarchies and Genome System Architecture

As an information storage medium for cellular computing, DNA is part of a living, reproducing, and evolving system. Consequently, DNA has to fulfill many functional requirements, some of them similar to those of electronic data storage systems and some of them quite different (e.g., von Neumann and Burks 1966). The list of genome functions includes the following:

- DNA condensation within the spatial confines of the nucleus or nucleoid.
- Transcriptional access to particular RNA and protein data files under appropriate circumstances.
- Maintenance of differentiated cellular states.
- Genome replication.
- Accurate transmission of genome copies to daughter cells.
- Proofreading and damage repair.
- DNA restructuring during the normal life cycle.
- DNA restructuring in response to crisis.

Without effective genome packaging, replication, transmission, and repair, no cell-based life form could reproduce reliably. Without DNA restructuring, no organism could evolve. These basic functions depend upon multiple features of genome organization.

2.5.1. Generic formatting signals and repetitive motifs

Each of the genome's functions requires it to be marked, or formatted, for specific interactions with particular cellular "machines" comprised of other molecules and supermolecular structures (e.g., the cell envelope in prokaryotes or the nuclear lamina and mitotic spindle in eukaryotes). This formatting requires a set of generic sequence codes that may be used multiple times at different places in the genome, such as replication and transcription start signals. The simpler the informational content of these codes is, the more efficiently they can be recognized by the appropriate cellular machinery. This provides pressure for a limited number of repetitive signals distributed throughout the genome (Shapiro and von Sternberg 2005). However, great specificity is also necessary, as in regulating the expression of thousands of data files in complex ways throughout cell and life cycles. Specificity is achieved through cooperative interactions between iterated and combined copies of the different formatting signals. Arranged in distinct combinations, the aggregated signals provide the structural basis for nucleoprotein complexes with the right computational properties (Davidson 2001). The use of iteration to facilitate cooperativity and combinatorial complexity provides additional pressure for repetition of generic formatting motifs. In other publications, different classes of functional repetitive motifs have been tabulated (Shapiro and von Sternberg 2005; Marino-Ramirez et al. 2005).

2.5.2. Different levels of genome organization

Molecular genetics and genome sequencing have taught us that genomes are organized for hierarchic regulation of data file expression. This realization dates back to the earliest studies on the control of bacterial protein synthesis, when it was recognized that expression of multiple loci could be regulated in a coordinated fashion (Jacob and Monod 1961; Ptashne 1986). Distinct data files can be coordinately controlled in two ways. One way is for the same sets of formatting motifs to be adjacent to unlinked coding sequences so that dispersed genetic loci can respond to a common transcription factor or set of factors. In this way, the repeated motifs serve as a physical basis for genome integration (Davidson and Britten 1979). Another form of hierarchical organization is for coordinately regulated data files to sit near each other along the DNA, where they can be transcribed from a single promoter or "indexed" into a single domain of differentially activated chromatin (Jenuwein 2002). Epigenetic control via chromatin formatting thus becomes a higher order form of regulation (van Driel et al. 2003; Kosak and Groudine 2004). This type of position-linked coordinate regulation probably underpins the conservation of large syntenic regions (segments containing a series of homologous genetic loci in identical order) in higher eukaryote chromosome evolution because it is a useful mechanism for closing down large genomic regions and maintaining states of cellular differentiation (Eichler and Sankoff 2003).

2.5.3. Evolution and reuse of genomic subsystems The emergence of repeated DNA structures comprising multiple formatting motifs (and sometimes coding sequences as well) is a reflection of the tendency for genomes to evolve by reusing integrated systems composed of various control circuit elements. Some of these structures, like the HOX complexes that execute segmental control in animal body plan development, extend over hundreds of thousands of base pairs and incorporate intricately arranged control modules for transcriptional regulation and chromatin configurations (Carroll 1995). These large complexes have been iterated and protected from disruption during animal evolution (Patel and Prince 2000; International Human Genome Sequencing Consortium 2001). Virtually all genomes also contain smaller DNA repeats, on the order of a few hundred to several thousand base-pairs, that contain assemblages of motifs regulating transcription and chromatin organization (Zhi et al. 2006). Many of these repeats are mobile genetic elements (MGEs) that have the capability to move to new locations and thus place different genetic loci under a common set of controls. For example, about 18–20% of the human genome is composed of dispersed LINE elements (International Human Genome Sequencing Consortium 2001), which contain a constellation of functional signals: well-documented promoters, enhancers, transcript elongation attenuators and nuclear matrix attachment regions as well as putative determinants for chromatin silencing (see Shapiro and von Sternberg [2005] for references). There is growing evidence that such mobile control modules have played an important role in the evolution of genomic regulatory hierarchies that operate on both the rapid computational and longer term epigenetic time scales (Britten 1996; Brosius 2003; Jordan et al. 2003; Peaston et al. 2004; Lippman et al. 2004).

2.5.4. Genome system architecture Genomes that are formatted and organized hierarchically for replication, transmission, regulated data file access, repair, and restructuring can be said to have a "genome system architecture," in much the same way that computer information storage and retrieval systems have system architectures independent of data file content (Shapiro 1999, 2002a, 2005; Shapiro and von Sternberg 2005). In genomes and computer systems, different architectures can achieve the same functions in different ways. We know that genome architecture can influence the expression and function of data file information without altering the data files themselves. The evidence is found in extensive documentation of "position effects," chromosome rearrangements or transpositions of intact genetic loci that alter regulation and phenotype (Spofford 1976; Hazelrigg et al. 1984; Levis et al. 1985; Schotta et al. 2003).

Aspects of system architecture that affect replication and chromosome transmission can influence reproductive compatibility without any change in somatic data file content. Hence, genomic steps toward speciation can occur before there is any alteration in adaptive phenotypes. Well-documented examples of architectural changes that affect germ line function and compatibility but not somatic phenotype include the presence of active transposons and retrotransposons (Bregliano and Kidwell 1983) as well as chromosome fusions (Hartmann and Scherthan 2004). It is likely that similar phylogenetic separation can result from chromosome inversions, distinct pericentromeric tandem repeat arrays, and amplification of different families of nonautonomous transposons and retrotransposons because these genomic features distinguish taxa that are otherwise quite similar (Tonzetich et al. 1988; Navarro and Barton 2003; Hey 2003; von Sternberg and Shapiro 2005; Shapiro and von Sternberg 2005).

A particularly interesting application of the genome system architecture concept occurs in prokaryotes, where lateral DNA transfer is widespread (Bapteste et al. 2004). Sequences encoding metabolic and ecological functions are easily exchanged between species. Thus, similar data files have been adapted to distinct system architectures. But integration of regulatory signals with the cellular hardware for protein synthesis means that coding sequences for molecules involved in transcription and translation cannot be exchanged without disrupting expression of proteins needed for virtually every cell function. This integration thus preserves the core genomic architecture of bacteria and archaea, each with a distinct characteristic transcriptional and translational system (Woese 2004).

3. Integration of the Genome into Distributed Cellular Information Processing

Conventional concepts postulate a kind of Cartesian distinction between genomic information stored in nucleic acids and executive function housed largely in proteins (Crick 1970). This dualistic view of how the genome operates in a cellular context has been invalidated by over four decades of research on the control of protein synthesis, dating back to pioneering work on the lac operon, and more recent studies of the role of signal transduction networks in regulating all aspects of genome function. In addition, contemporary cell biology has revealed major realms of information processing that do not directly involve DNA.

3.1. Cellular Information Processing That Does Not Involve the Genome

Studies of many processes, from metabolic pathways to cell migration, have revealed signal transduction systems that operate computationally without involving the genome. The control of bacterial swimming by the chemotaxis control circuit is a basic paradigm for these extragenomic networks (Szurmant and Ordal 2004; Armitage et al. 2005). Other well-understood examples include rapid control of catabolism and biosynthesis, aggregation of surface receptors in response to ligands (Wulfing et al. 2002; Bray and Duke 2004; Murai and Pasquale 2004), protein and vesicle targeting to distinct compartments (Bonifacino and Glick 2004; Pool 2005), endocytosis (Neel et al. 2005; Stuart and Ezekowitz 2005) and cytoskeletal reorganization (Pollard and Borisy 2003; Pelkmans 2005). Clearly, cell computations guiding important processes can occur without accessing DNA data files.

3.2. Involvement of the Whole Cell in Computations Involving the Genome

Our understanding of how cells compute using the genome depends upon study of model systems, starting with the *lac* operon and bacteriophage λ (Ptashne 1986) and continuing through to multicellular development in *Drosophila* and other model organisms, such as sea urchins (Davidson 2001). In all these systems, there is communication between nuclear or DNA-binding transcription factors and molecules in other compartments of the cell.

The *lac* operon presents the simplest and most thoroughly analyzed case. In *lac* regulation, metabolic interactions intervene at two important places (Box 2). Because *lac* derepression only occurs with the participation of cytoplasmic enzymes and membrane transport proteins, one cannot make a basic distinction between functional metabolism and information processing nor can one model *lac* operon control as a function solely of transcription factors.

Box 2. Nongenomic components of *lac* regulation. The availability of lactose as a substrate is indicated through its conversion to allolactose inducer; a process that requires background levels of cytoplasmic LacZ (beta-galactosidase) and membrane-bound LacY (lactose permease). Without lactose transport and enzymatic conversion to allolactose, the operon cannot be derepressed. The second metabolic intervention comes through the glucose-specific EnzII^{GLC} membrane-associated component of the PTS transport system, which serves as a sensor for external glucose and as a regulator of cAMP synthesis. Only when external glucose is absent do the cells contain the high levels of cAMP needed for normal transcription of the *lac* operon. Readers unfamiliar with the lac operon can best inform themselves through a simple internet search that will turn up several basic descriptions of the system and consult Shapiro (2002) for a more detailed summary and references.

Analogous involvement of extragenomic processes occurs in eukaryotic signal transduction (Gerhart and Kirschner 1997; Alberts et al. 2002). Control of genome transcription by environmental conditions, nutrition, physiology, pheromones, hormones, intercellular signaling, cell injury, or checkpoints is invariably subject to extragenomic inputs. Take transcription as a case in point. Transcription factors can be modified by protein kinases and phosphatases, which are often linked to cell-surface receptors, receptor-activated G proteins, or second messengers. Transcription factor persistence can be controlled by ubiquitinating activities, transcription factor localization by connection to large cellular complexes, and transcription factor function by the presence or absence of inhibitory and activating ligands. In other words, the transcriptional control circuitry of every cell is in continuous communication with the rest of that cell.

3.3. The Conceptual Significance of Communication with Extragenomic Signal Transduction

The “filtering” of genomic regulatory information through communication with cytoplasmic, organellar and surface compartments has fundamental consequences for understanding genome function:

- It explains how and why organismal phenotypes are not hard-wired in the genome.
- It indicates that attempts to portray cell regulatory systems as direct “gene to gene” circuits (“gene networks”) are not realistic.
- It allows us to understand how checkpoints and other feedback mechanisms can modulate control circuits in response to unpredictable events. This means that cells can adjust to normal function and phenotype despite genetic deficiencies, developmental errors, or experimental disruptions. Because

control regimes are distributed and computational, it appears that attempts to pin down the molecular nature of many classical developmental regulatory functions, such as “Spemann’s organizer” (Spemann and Mangold 1924; Spemann 1938), are destined to be frustrated because of the distributed, computational nature of the responsible control regimes.

4. The Genome as a Read-Write Information Storage System

It is obvious in electronic information processing that a RW memory is far more useful than a ROM memory. The ability to add and subtract applications and modify data files endows the whole system with adaptability to many different tasks and extends its effective lifetime for many years as software evolves. Can we see parallel RW capacities in cellular information processing?

4.1. Computational and Epigenetic Storage as Read-Write Memory

It is not difficult to see how short- and medium-term information is written into the genome. The information stored in nucleoprotein complexes, chromatin domains, and chemical modifications of DNA is written by cell computing functions and serves as the basis for subsequent computations. Changes in chromatin configurations that do not alter DNA sequence content can be perpetuated over cell or organismal generations (Jaenisch and Bird 2003; Box 3). Somatically heritable chromatin structures appear to serve as one mechanism for the maintenance of differentiated cell states (Gurdon et al. 2003). Accordingly, basic processes like cellular differentiation may usefully be conceptualized in terms of RW memory.

Box 3. Epigenetic storage. Perhaps the most distinctive example of cell-determined heritable epigenetic information is the phenomenon of imprinting: the expression of certain genetic loci is determined by the sex of the parent from whom they were inherited (Mann 2001; Li 2002). Such loci are “imprinted” during spermatogenesis and oogenesis so that the cells of the resulting zygote can distinguish them and express only the information contained in the paternally or maternally inherited copy. In the next generation, the imprinting can change according to the sex of the individual.

4.2. Natural Genetic Engineering Tools That Alter DNA Sequence Information

The RW aspect of information stored in DNA sequences is harder to see because it has been assumed for so long that this information changes randomly and accidentally. Nonetheless, a major lesson of the last half-century of molecular genetics is the ubiquity of cellular biochemical activities that have the capacity to change sequence information in DNA molecules (Box 4). In other words, we now understand in considerable detail the biochemical machinery cells have available to write new sequence information. Like all cellular biochemistry, the molecules and complexes that generate novel DNA structures are subject to control by signal transduction networks and are activated in response to particular stimuli (McClintock 1984; Wessler 1996; Shapiro 1997). Of equal importance is the growing body of data indicating that natural genetic engineering tools can be targeted to regions, sites, or specific internucleotide bonds in the genome. The chief molecular mechanisms for targeting natural genetic engineering functions are protein and RNA recognition of specific nucleotide sequences and coupling of DNA rearrangement functions to transcriptional control or chromatin-formatting functions (Box 5). Thus, the molecular basis exists for a biologically regulated process of generating novel DNA sequence information.

Box 4. Natural genetic engineering activities. They include nucleases, ligases, polymerases (especially mutator polymerases), homologous recombination proteins, nonhomologous end-joining systems, site-specific recombination systems, DNA transposons, reverse transcriptases and retrotransposons, and combinations of all the above (Craig et al. 2002). These natural genetic engineering tools can generate a wide variety of novel sequence structures that include single nucleotide alterations, novel untemplated oligonucleotides, cDNA copies of processed RNA molecules, duplication and insertion of segments ranging in size from dozens to millions of base-pairs in length, and rearrangements at all length scales of existing DNA segments (inversions, deletions, translocations, etc.).

4.3. Natural Genetic Engineering in Regular Organismal Life Cycles

In many organisms, controlled DNA rearrangement is part of the normal life cycle. Since the examples include bacterial phase variation and cell differentiation (van der Woude and Baumler 2004), yeast mating-type interconversion (Haber 1998), macronuclear development in ciliated protozoa (Prescott 2000), chromatin diminution in invertebrate somatic development (Muller et al. 1996; Reddi et al. 2001; Goday and Esteban 2001), and vertebrate immune system rearrangements, organisms utilizing DNA restructuring are far more taxonomically diverse than generally appreciated. In a number of these highly evolved instances of regular natural genetic engineering, we know that the DNA rearrangements are tightly regulated with respect to when in the life cycle, in which cells, and where in the genome they occur (Box 6). These highly evolved cases, where DNA rearrangements fulfill specific

Box 5. Nonrandomness and targeting in natural genetic engineering. Although published articles (and unpublished referee reports) frequently assert that insertions of mobile genetic elements and sites of action of other DNA rearrangement systems are random in the genome, the evidence for targeting is quite extensive (specific references in Table 1, Shapiro 2005).

Elucidated molecular targeting mechanisms include the following:

- Sequence recognition by proteins (yeast mating-type switching, ribosomal LINE element insertions, group I homing introns, VDJ joining).
- Protein–protein interaction (Ty retrotransposon targeting).
- Sequence recognition by RNA (reverse splicing of group II retrohoming introns).
- Transcriptional activation (somatic hypermutation and class-switch recombination).

Well-documented targeting phenomena whose mechanisms remain to be determined include the following:

- Telomere targeting of certain LINE elements in insects.
- HIV and MLV retrovirus targeting upstream of transcribed regions in the human and mouse genomes.
- P factor "homing" directed by internal transcription factor sites and chromatin signals.

Box 6. Tight regulation of programmed DNA rearrangements. In ciliated protozoa, the germline genome is quite literally chopped into hundreds of thousands of fragments and a new somatic genome is efficiently reconstructed from a subset of the fragments, many of which need to be connected in new orders to reconstitute functional coding sequences (Prescott 2000). Even in the immune system, where strict determinism would be unproductive in making binding proteins for unpredictable invaders, the locations of VDJ rearrangements are precisely targeted by special sequences so that the antigen-binding region of the antibody or T-cell receptor molecules is specifically diversified. The locations of further immune system rearrangements generating distinct immunoglobulin classes are determined by lymphokine signaling molecules that activate transcription at the sites of chromosome breakage and rejoining (Kinoshita and Honjo 2001; Gellert 2002).

purposes in the life cycle, serve as important counterexamples to the widespread belief that genomic changes must occur stochastically and cannot be targeted in any functional way. These examples of programmed natural genetic engineering are either reversible, as in vegetative microbes like bacteria (van der Woude and Baumler 2004) and yeast (Haber 1998), or they occur in cells and nuclei that do not contribute to the germline; hence they do not violate the theoretical principles established for self-replicating systems by von Neumann (von Neumann and Burks 1966).

4.4. Documented and Potential Evolutionary Roles of Natural Genetic Engineering

Whole genome sequencing has led to a number of discoveries that are challenging both for traditional "gene-based" ideas of genome content and for random mutation models of genome change. These discoveries include the following:

- The relatively small number of genetic loci in the human and other higher metazoan genomes.
- The iteration of protein-coding determinants to generate taxonomically specific paralogue families in all organisms, from bacteria to mammals (Jordan et al. 2001; Nei and Rooney 2005).
- The surprisingly high abundance of dispersed repetitive sequence elements in genomes (Shapiro and von Sternberg 2005), even in compressed genomes of rapidly growing organisms like bacteria (Shapiro 2002c).
- Taxonomic specificity of repeat element families, such as mammalian SINEs (von Sternberg and Shapiro 2005).
- The abundance of segmental duplications in eukaryotic genomes, with duplications ranging in size from individual exons to long syntenic regions carrying dozens of genetic loci (Arabidopsis Genome Initiative 2000; Eichler 2001; International Human Genome Sequencing Consortium 2001).
- Conservation and scrambling of syntenic regions in the genomes of organisms as distantly related as primates and rodents (Mouse Genome Sequencing Consortium 2002) or mustards and cereals (Goff et al. 2002).

These and other important features of the whole genome structure indicate that evolutionary changes involve multiplication of mobile repeats, coding sequence duplication and transposition, and chromosome breakage and rejoining in new combinations. In other words, the rearrangements that result from action by natural genetic engineering systems become glaringly apparent when sequenced genomes are compared. In some cases, we can be quite confident that changes resulted from the activity of a specific natural genetic engineering system (Box 7). Because we can conclude from whole genome sequences that natural genetic engineering functions have played major roles in genome evolution, it is logical to postulate that the informatic/computational inputs known to be associated with natural genetic engineering systems may well have participated in influencing the novel genome configurations that were subsequently tested by selection. Thus, instead of thinking about evolutionary change as an adaptively blind process, it may be radically reconceptualized as a computationally guided example of system engineering.

There are several advantages to an engineered process of evolution. Natural genetic engineering can increase the

Box 7. Two examples illustrate the capacity of natural genetic engineering systems to generate coding sequence duplications, a fundamental process in genome evolution.

(i) Mammalian species differ in their repertoires of olfactory receptors, which comprise a family of amplified proteins. Many of the coding sequences for proteins in this family lack introns and therefore must have arisen as "retrogenes" generated by LINE element reverse transcriptase functions (Brosius 1999). (ii) There are thousands of segmental duplications in the rice genome and at least 3000 of these are inside so-called Pack-MULE DNA transposons incorporating exons from other genomic locations regions (Jiang et al. 2004). Thus, both protein amplification and exon shuffling in rice evolution were clearly mediated by DNA transposition functions.

efficiency of searching for genome configurations that encode functional complex systems and can favor the elaboration of hierarchic system architectures. It reduces the degrees of freedom for searches through genome space from virtual infinity (for random changes) to large but much smaller numbers (for specific kinds of DNA rearrangements). At the same time as reducing the search space, natural genetic engineering often produces just the kinds of genomic changes that are most likely to prove adaptive. For example, by exon shuffling, natural genetic engineering inserts into a genetic locus a DNA segment encoding one or more already functional domains, which are far more likely to add new capabilities to a protein than are random sequence variants or the addition of random polypeptide segments. Similarly, insertion of a mobile element carrying an integrated package of transcription and chromatin-formatting signals can place existing coding regions under novel controls so that established functional products can be expressed under conditions where they were previously absent. Such processes have been well documented in the laboratory for decades and are copied in human genetic engineering, where swapping control regions and coding sequences for protein domains is common practice. Moreover, the evidence is quite good (and continually growing stronger) that these processes have occurred during the evolution of sequenced genomes (Britten 1996; International Human Genome Sequencing Consortium 2001; Brosius 2003; Jordan et al. 2003; Peaston et al. 2004).

By making sure that genomes in normally reproducing organisms are stable and that the genomes of cells under stress are mutable, computational networks regulating natural genetic engineering functions provide hereditary variability when it is most needed (McClintock 1984; Wessler 1996; Shapiro 1997). Targeting of natural genetic engineering functions can limit change to regions where it has the highest probability of being functional (Box 5). For example, the bias for many retrotransposon insertions to occur upstream of transcription start sites (Table 1 of Shapiro 2005) prevents damage to functional coding elements and enhances the potential for a constructive regulatory change. It is relevant that such upstream retrotransposon insertions are the most common mutations found in budding yeast after selection for increased protein expression (Errede et al. 1981) and are a mechanism for retroviral oncogenesis (i.e., for initiating tumor cell evolution; Butturini et al. 1988). Targeting processes can also facilitate fine-tuning of individual components (microevolution) after initial rearrangements establish a new system (in a way that is reminiscent of progress in human engineering). The immune system provides an instructive example. "Rearrangement followed by fine-tuning" occurs when exon joining and clonal selection are followed by somatic hypermutation targeted to the exons encoding antigen-binding domains (Gellert 2002).

The action of natural genetic engineering systems imparts structural characteristics to genomes consistent with whole genome sequencing results. By mediating duplications and rearrangements of DNA segments ranging in size from a few hundred to several million base-pairs (Harden and Ashburner 1990; Moran et al. 1999; Bailey et al. 2003), natural genetic engineering facilitates the establishment and amplification of higher order genomic subsystems, such as homeodomain complexes (Carroll 1995; Patel and Prince 2000) and large syntenic regions. The tendency to amplify progressively larger subsystems may help account for the hierarchic nature of genome coding (van Driel et al. 2003). Change by natural genetic engineering also leads to the accumulation of dispersed repeats. Since dispersed repeats influence both coding sequence expression and physical organization of genomes, it is reasonable to entertain the functionalist hypothesis that repeat accumulation represents the establishment of a system architecture required for effectively integrated genome functioning (Shapiro and von Sternberg 2005).

5. What Is Fundamentally New in This View of Genome Informatics?

Conventional theories of genetics and evolution were formulated before the demonstration that DNA carries hereditary information (Avery et al. 1944) or the elucidation of the double helix and A-T/G-C base-pairing (Watson and Crick 1953). Since 1953, a major emphasis of molecular genetics has been to reframe conventional genetic concepts in terms of the chemical structure of DNA. Although prodigiously useful in terms of technology development, such "classical" molecular genetics has not proved particularly helpful in producing a conceptual framework for interpreting the expanding catalogue of "genes," proteins, RNAs, pathways, and networks that molecular cell biology and genome sequencing have uncovered (Alberts et al. 2002). Many authors have suggested a more computational and integrated approach (Bray 1990, 1995; Gerhart and Kirschner 1997; Hartwell et al. 1999). In this essay, an explicitly informatic approach has been embraced to stimulate the formulation of a more consistently computational way of thinking about genome function. There are four conceptual novelties to this approach.

5.1. There Are No Fundamental Genomic Units, Only Systems

A major (but often unstated) goal of conventional approaches to formulating general theories of genome organization and function has been to identify basic "units," such as genes. Molecular analysis has revealed that there are no indivisible units in the genome. Coding sequences, regulatory signals, genetic loci, and structural domains (centromeres, telomeres, etc.) are subject to deconstruction into smaller components as well as to combinatorial modification by rearrangement of those components. By emphasizing the systemic nature of genome function, genome informatics avoids the reductionist fallacy of claiming that a given segment of the genome determines a particular trait. In the computational view, each data file or repetitive signal may contribute a necessary component to phenotypic expression, but individual sequence elements can never be sufficient to encode a trait by themselves. As a reviewer of this paper expressed it, the computational view "allows us to think of functional units as distributed over the genome and linked by a set of functional regulators, just as data files are not necessarily stored contiguously on a computer disk but are linked together in a look-up table." By directing attention away from single genetic loci in phenotype determination, the systemic perspective helps resolve an apparent paradox in evolutionary studies: how conserved regulatory and morphogenetic functions have come to encode such diverse organismal phenotypes (Duboule and Wilkins 1998). If the phenotype results from integrated action of many genomic elements, then various combinations of basic elements can produce a wide variety of complex characters in a way that is analogous to building different electronic circuits out of the same array of resistors, capacitors, transistors, and other components. Formal representation of genomic networks will have to incorporate this intrinsically systemic view.

5.2. Genome Information Transfer Is an Obligate Multidirectional Process

Genome informatics postulates that information flows both to and from DNA. Since DNA is inert outside its cellular context, the genome only functions in constant communication with other molecules in the cell. In essence, functional models will be circuit-like in structure with provision for modulating inputs at each genomic node. Thus, how genome information is expressed is necessarily a function of cell history. So it is not difficult to see that one genome can contain the information specifying different kinds of cells (as in cellular differentiation) and different kinds of organisms (as in complex life cycles). Because there is no isolated Cartesian information-processing compartment in the cell, there can be no one-way information transfer of instructions from the genome for execution in "noninformatic" compartments, and thus DNA sequence information cannot determine phenotypes in a hard-wired fashion.

5.3. So-Called Noncoding Genome Information Is Essential

Genome informatics predicts major functional roles for repetitive DNA sequence elements that format and organize the genome and its data files for the multiple tasks that need to be accomplished during cell and organismal life cycles (Shapiro and von Sternberg 2005). These repetitive, so-called noncoding DNA elements in fact constitute a variety of distinct codes influencing genome packaging, expression, maintenance, repair, and restructuring. Several formatting signals (oligonucleotide motifs) often aggregate into larger composite elements, the dispersed repeats, that provide defined constellations of coding elements at multiple locations across the genome (Zhi et al. 2006). Combinations of repeats and data files arranged in larger chromosome domains define a characteristic genome system architecture for each taxonomic group, with structural and repetitive elements making fundamental contributions to organismal phenotypes.

5.4. Genome Function in Heredity and Evolution Is Inherently Computational

The obligatory dependence of genome function on other cellular molecules (Section 2.3) connects the genome to cellwide signaling and computing networks. The molecules involved in chromatin formation, transcription, replication, chromosome movement, repair and DNA restructuring respond to multiple intra- and extracellular signals. How these molecules complex with and operate on the genome is, in turn, a function of genome sequence, existing molecular modifications of the DNA (covalent and noncovalent), and localization of the genome within the nucleus or nucleoid. In other words, genome function results from an intrinsically complex series of molecular interactions. These interactions involve sequence arrangements, epigenetic modifications, and transient nucleoprotein complexes, thereby integrating all three levels of stored information (Sections 2.2.1–2.2.3) with extragenomic cellular inputs. We know that these complex interactions produce outputs with remarkable accuracy and coordination because unbalanced growth, chromosome nondisjunction, inappropriate cell differentiation, and unprogrammed cell death are quite rare in organisms growing under normal conditions (when measured, far less than 1%; e.g., Hall et al. 2003). This indicates that there is logical (computational) guidance for the huge number of biochemical and biomechanical operations needed for every cell cycle. Without such guidance, cellular complexity would lead to chaotic transformations and trapping in a limited number of "attractor" states, as occurs in complex dynamic systems lacking computational feedback.[1] We are beginning to acquire knowledge of how these guiding computations operate (e.g., checkpoints), but elucidating their underlying general principles and mechanisms is the key research goal for this new century.

Finally, we know that evolution has produced these incredibly efficient expert cellular systems, and whole genome sequences indicate to us that (computationally regulated) natural genetic engineering functions have been important in genome evolution. Thus, another major research goal is to investigate the role of cellular computation in genome evolution itself (McClintock 1984). As outlined earlier (Section 4.4), the role of nonrandom natural genetic engineering processes introduces major advantages to plausible theories of evolutionary change. Adding computational guidance of these processes, we may now begin to think of evolution in terms of systems engineering rather than as a blind walk through the thickets of purifying selection.

Note

[1] See http://www.eecs.berkeley.edu/∼lieber/Textbook1.html.

References

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) Molecular Biology of the Cell. 4th ed. New York: Garland.

Almeida R, Allshire RC (2005) RNA silencing and genome regulation. Trends in Cell Biology 15: 251–258.

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

Armitage JP, Holland IB, Jenal U, Kenny B (2005) "Neural networks" in bacteria: Making connections. Journal of Bacteriology 187: 26–36.

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

Atlan H, Koppel M (1990) The cellular computer DNA: Program or data. Bulletin of Mathematical Biology 52: 335–348.

Avery OT, MacLeod CM, McCarty M (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a deoxyribonucleic acid fraction isolated from Pneumococcus Type III. Journal of Experimental Medicine 79: 137–158.

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

Bapteste E, Boucher Y, Leigh J, Doolittle WF (2004) Phylogenetic reconstruction and lateral gene transfer. Trends in Microbiology 12: 406–411.

Benzer S (1962) The fine structure of the gene. Scientific American 206: 70–84.

Bernstein E, Allis CD (2005) RNA meets chromatin. Genes and Development 19: 1635–1655.

Bibikova M, et al. (2006) Human embryonic stem cells have a unique epigenetic signature. Genome Research 16: 1075–83.

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

Bonifacino JS, Glick BS (2004) The mechanisms of vesicle budding and fusion. Cell 116: 153–166.

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.

Bray D, Duke T (2004) Conformational spread: The propagation of allosteric states in large multiprotein complexes. Annual Review of Biophysics and Biomolecular Structure 33: 53–73.

Bregliano, JC, Kidwell MG (1983) Hybrid dysgenesis determinants: In: Mobile Genetic Elements (Shapiro JA, ed), 363–410. New York: Academic Press.

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

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

Brosius J (2003) The contribution of RNAs and retroposition to evolutionary novelties. Genetica 118: 99–116.

Butturini A, Sthivelman E, Canaani E, Gale RP (1988) Oncogenes in human leukemias. Cancer Investigation 6: 305–316.

Carroll SB (1995) Homeotic genes and the evolution of arthropods and chordates. Nature 376: 479–485.

Cases I, de Lorenzo V (1998) Expression systems and physiological control of promoter activity in bacteria. Current Opinion in Microbiology 1: 303–310.

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

Crick F (1970) Central dogma of molecular biology. Nature 227: 561–563.

Davidson EH (2001) Genomic Regulatory Systems: Development and Evolution. San Diego: Academic Press.

Davidson EH, Britten RJ (1979) Regulation of gene expression: Possible role of repetitive sequences. Science 204: 1052–1059.

Driesch H (1908) The Science and Philosophy of the Organism. London: A& C Black.

Duboule D, 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, Sankoff D (2003) Structural dynamics of eukaryotic chromosome evolution. Science 301: 793–797.

Errede B, Cardillo TS, Wever G, Sherman F, Stiles JI, Friedman LR, Sherman F (1981) Studies on transposable elements in yeast. I. ROAM mutations causing increased expression of yeast genes: Their activation by signals directed toward conjugation functions and their formation by insertion of Ty1 repetitive elements. II. deletions, duplications, and transpositions of the COR segment that encompasses the structural gene of yeast iso-1-cytochrome c. Cold Spring Harbor Symposium Quantitative Biology 45 (Pt 2): 593–607.

Galli C, Lagutina I, Lazzari G (2003) Introduction to cloning by nuclear transplantation. Cloning Stem Cells 5: 223–232.

Gao S, Latham KE (2004) Maternal and environmental factors in early cloned embryo development. Cytogenetic and Genome Research 105: 279–284.

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

Gerhart J, Kirschner M (1997) Cells, Embryos, and Evolution. Malden, MA: Blackwell Science.

Goday C, Esteban MR (2001) Chromosome elimination in sciarid flies. Bioessays 23: 242–250.

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, and 45 others (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science 296: 92–100.

Gurdon JB, Byrne JA, Simonsson S (2003) Nuclear reprogramming and stem cell creation. Proceedings of the National Academy of Sciences USA 100 (Suppl 1): 11819–11822.

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

Hall IM, Noma K, Grewal SI (2003) RNA interference machinery regulates chromosome dynamics during mitosis and meiosis in fission yeast. Proceedings of the National Academy of Sciences USA 100: 193–198.

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

Hartmann N, Scherthan H (2004) Characterization of ancestral chromosome fusion points in the Indian muntjac deer. Chromosoma 112: 213–220.

Hartwell L (1992) Defects in a cell cycle checkpoint may be responsible for the genomic instability of cancer cells. Cell 71: 543–546.

Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402 (Suppl 6761): C47–C52.

Hazelrigg T, Levis R, Rubin GM (1984) Transformation of white locus DNA in drosophila: Dosage compensation, zeste interaction, and position effects. Cell 36: 469–481.

Henikoff S, Ahmad K (2005) Assembly of variant histones into chromatin. Annual Review of Cell and Developmental Biology 21: 133–153.

Hey J (2003) Speciation and inversions: Chimps and humans. Bioessays 25: 825–828.

International Human Genome Sequencing 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. Journal of Molecular Biology 3: 318–356.

Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. Nature Genetics 33 (Suppl): 245–254.

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

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature 431: 569–573.

Jordan IK, Makarova KS, Spouge JL, Wolf YI, Koonin EV (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. Genome Research 11: 55–65.

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

Judson HF (1979) The Eighth Day of Creation: Makers of the Revolution in Biology. New York: Simon and Schuster.

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

Kosak ST, Groudine M (2004) Gene order and dynamic domains. Science 306: 644–647.

Kuhn TS (1962) The Structure of Scientific Revolutions. Chicago: University of Chicago Press.

Kunkel TA, Erie DA (2005) DNA mismatch repair. Annual Review of Biochemistry 74: 681–710.

Landy A (1989) Dynamic, structural, and regulatory aspects of lambda site-specific recombination. Annual Review of Biochemistry 58: 913–949.

Levis R, Hazelrigg T, Rubin GM (1985) Effects of genomic position on the expression of transduced copies of the white gene of Drosophila. Science 229: 558–561.

Li E (2002) Chromatin modification and epigenetic reprogramming in mammalian development. Nature Reviews Genetics 3: 662–673.

Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, Carrington JC, Doerge RW, Colot V, Martienssen R (2004) Role of transposable elements in heterochromatin and epigenetic control. Nature 430: 471–476.

London M, Hausser M (2005) Dendritic computation. Annual Review of Neuroscience 28: 503–532.

Lucas I, Hyrien O (2000) Hemicatenanes form upon inhibition of DNA replication. Nucleic Acids Research 28: 2187–2193.

Mann JR (2001) Imprinting in the germ line. Stem Cells 19: 287–294.

Marino-Ramirez L, Lewis KC, Landsman D, Jordan IK (2005) Transposable elements donate lineage-specific regulatory sequences to host genomes. Cytogenetic and Genome Research 110: 333–341.

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

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

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

Murai KK, Pasquale EB (2004) Eph receptors, ephrins, and synaptic function. Neuroscientist 10: 304–314.

Muller C, Leutz A (2001) Chromatin remodeling in development and differentiation. Current Opinion in Genetics and Development 11: 167–174.

Muller F, Bernard V, Tobler H (1996) Chromatin diminution in nematodes. Bioessays 18: 133–138.

Muotri AR, Chu VT, Marchetto MCN, Deng W, Moran JV, Gage FH (2005) Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435: 903–910.

Murray AW (2004) Recycling the cell cycle: Cyclins revisited. Cell 116: 221–234.

Navarro A, Barton NH (2003) Chromosomal speciation and molecular divergence: accelerated evolution in rearranged chromosomes. Science 300: 321–324.

Neel NF, Schutyser E, Sai J, Fan GH, Richmond A (2005) Chemokine receptor internalization and intracellular trafficking. Cytokine and Growth Factor Reviews 16: 637–658.

Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annual Review of Genetics 22: 121–152.

Nusslein-Volhard C (1991) Determination of the embryonic axes of Drosophila. Development, (Suppl 1): 1–10.

Patel NH, Prince VE (2000) Beyond the Hox complex. Genome Biology 1: reviews 1027.1–1027.4.

Peaston AE, et al. (2004) Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Developmental Cell 7: 597–606.

Pelkmans L (2005) Viruses as probes for systems analysis of cellular signalling, cytoskeleton reorganization and endocytosis. Current Opinion in Microbiology 8: 331–337.

Pollard TD, Borisy GG (2003) Cellular motility driven by assembly and disassembly of actin filaments. Cell 112: 453–465.

Pool MR (2005) Signal recognition particles in chloroplasts, bacteria, yeast and mammals (review). Molecular Membrane Biology 22: 3–15.

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

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

Razin SV, Petrov A, Hair A, Vassetzky YS (2004) Chromatin domains and territories: Flexibly rigid. Critical Reviews in Eukaryotic Gene Expression 214: 79–88.

Rothenburg S, Koch-Nolte F, Haag F (2001) DNA methylation and Z-DNA formation as mediators of quantitative differences in the expression of alleles. Immunological Reviews 184: 286–298.

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

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: Molecular Strategies for Biological Evolution (Caporale L, ed), Annals of the New York Academy of Sciences 870: 23–35.

Shapiro JA (2002a) Genome organization and reorganization in evolution: Formatting for computation and function. In: From Epigenesis to Epigenetics: The Genome in Context (Van Speybroeck L, Van de Vijver G, De Waele D, eds), Annals of the New York Academy of Sciences 981: 111–134.

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

Shapiro JA (2002c) Repetitive DNA, genome system architecture and genome reorganization. Research in Microbiology 153: 447–453.

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

Shapiro JA, Dworkin M, eds (1997) Bacteria as Multicellular Organisms. New York: Oxford University Press.

Shapiro JA, von Sternberg R (2005) Why repetitive DNA is essential for genome function. Biological Review 80: 227–250.

Sidiropoulou K, Pissadaki EK, Poirazi P (2006) Inside the brain of a neuron. EMBO Reports 7(9): 886–892.

Smith GR (1994) Hotspots of homologous recombination. Experientia 50: 234–241.

Spemann H (1938) Embryonic Development and Induction. New Haven, CT: Yale University Press.

Spemann H, Mangold H (1924) Über Induktion von Embryonanlagen durch Implantation artfremder Organisatoren. Roux’ Archiv für Entwicklungsmechanik 100: 599–638.

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

Stros M, Muselikova-Polanska E, Pospisilova S, Strauss F (2004) High-affinity binding of tumor-suppressor protein p53 and HMGB1 to hemicatenated DNA loops. Biochemistry 43: 7215–7225.

Stuart LM, Ezekowitz RA (2005) Phagocytosis: Elegant complexity. Immunity 22: 539–550.

Szurmant H, Ordal GW (2004) Diversity in chemotaxis mechanisms among the bacteria and archaea. Microbiology and Molecular Biology Reviews 68: 301–319.

Tonzetich J, Lyttle TW, Carson HL (1988) Induced and natural break sites in the chromosomes of Hawaiian Drosophila. Proceedings of the National Academy of Sciences USA 85: 1717–1721.

Turing AM (1950) Computing machinery and intelligence. Mind 59: 433–460.

van der Woude MW, Baumler AJ. Phase and antigenic variation in bacteria. Clinical Microbiology Reviews 17: 581–611.

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

von Neumann J, Burks AW (1966) Theory of Self-Reproducing Automata. Urbana: University of Illinois Press.

von Sternberg R, Shapiro JA (2005) How repeated retroelements format genome function. Cytogenetic and Genome Research 110: 108–116.

Wang JY, Syvanen M (1992) DNA twist as a transcriptional sensor for environmental changes. Molecular Microbiology 6: 1861–1866.

Watson JD, Crick FHC (1953) Molecular structure of nucleic acids. Nature 171: 737–738.

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

Williamson JR (1994) G-quartet structures in telomeric DNA. Annual Review of Biophysics and Biomolecular Structure 23: 703–730.

Wilmut I, Paterson L (2003) Somatic cell nuclear transfer. Oncology Research 13: 303–307.

Woese CR (2004) A new biology for a new century. Microbiology and Molecular Biology Reviews 68: 173–186.

Wulfing C, Tskvitaria-Fuller I, Burroughs N, Sjaastad MD, Klem J, Schatzle JD (2002) Interface accumulation of receptor/ligand couples in lymphocyte activation: Methods, mechanisms, and significance. Immunological Reviews 189: 64–83.

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

Zaidi SK, Young DW, Choi JY, Pratap J, Javed A, Montecino M, Stein JL, van Wijnen AJ, Lian JB, Stein GS (2005) The dynamic organization of gene-regulatory machinery in nuclear microenvironments. EMBO Reports 6: 128–133.

Zhi D, Raphael B, Price A, Tang H, Pevzner P (2006) Identifying repeat domains in large genomes. Genome Biology 7: R7.