



Review

# How life changes itself: The Read–Write (RW) genome

James A. Shapiro

*Dept. of Biochemistry and Molecular Biology, University of Chicago, GCIS W123B, 979 E. 57th Street, Chicago, IL 60637, USA*

Received 27 June 2013; accepted 2 July 2013

Available online 8 July 2013

Communicated by M. Frank-Kamenetskii

## Abstract

The genome has traditionally been treated as a Read-Only Memory (ROM) subject to change by copying errors and accidents. In this review, I propose that we need to change that perspective and understand the genome as an intricately formatted Read–Write (RW) data storage system constantly subject to cellular modifications and inscriptions. Cells operate under changing conditions and are continually modifying themselves by genome inscriptions. These inscriptions occur over three distinct time-scales (cell reproduction, multicellular development and evolutionary change) and involve a variety of different processes at each time scale (forming nucleoprotein complexes, epigenetic formatting and changes in DNA sequence structure). Research dating back to the 1930s has shown that genetic change is the result of cell-mediated processes, not simply accidents or damage to the DNA. This cell-active view of genome change applies to all scales of DNA sequence variation, from point mutations to large-scale genome rearrangements and whole genome duplications (WGDs). This conceptual change to active cell inscriptions controlling RW genome functions has profound implications for all areas of the life sciences.

© 2013 Elsevier B.V. All rights reserved.

*Keywords:* Epigenetics; Natural genetic engineering (NGE); Mobile genetic elements (MGEs); Genome inscriptions

## Contents

0.	Introduction: Hereditary change – what have we learned in the last 60 years? . . . . .	288
1.	The genome is a formatted data storage system for living cells . . . . .	289
2.	Different types of genome inscriptions occur over three distinct biological time scales . . . . .	293
3.	Cells and their genomes are not passive victims of replication errors or DNA damage . . . . .	298
4.	Cell action in genome changes: Natural genetic engineering . . . . .	299
5.	NGE and mobile elements have played an important role in genome writing during the historical record . . . . .	308

*Abbreviations:* nc, non-coding; CDS, coding sequence; CRM, *cis*-regulatory module; TF, transcription factor; C, cytosine; U, uracil; A, adenine; I, inosine; G, guanine; T, thymine; RITS, RNA interference by transcriptional silencing; CRISPR, clustered regularly interspaced short palindromic repeats; WGD, whole genome duplication; MGE, mobile genetic element; TPRT, target-primed reverse transcription; LTR, long terminal repeat; DS, double-strand; CSR, class switch recombination; Rec, recombination; NHEJ, non-homologous end joining; NER, nucleotide excision repair; BER, base excision repair; NGE, natural genetic engineering; CNE, conserved nucleotide element; K, lysine; R, arginine; TIR, terminal inverted repeat; MB, mega-base-pairs.

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

*URLs:* <http://www.huffingtonpost.com/james-a-shapiro>, <http://shapiro.bsd.uchicago.edu>.

6. Conclusions and outstanding issues . . . . .	313
Acknowledgements . . . . .	316
References . . . . .	316

---

## 0. Introduction: Hereditary change – what have we learned in the last 60 years?

Understanding the proper role of the genome is a central issue in biology. The susceptibility of hereditary mechanisms to biological influences was a matter of intense debate in the late 18th and 19th centuries [1–3]. With the twin triumphs of Mechanism over Vitalism and Darwinism over Lamarckism in the first half of the 20th century, the Weismann idea of the segregated germ plasm directing the actions of the cell and organism in a hard-wired fashion has been the prevailing view. Ever since the formulation of the neo-Darwinist Modern Synthesis evolutionary theory in the 1930s and 1940s, it has been an article of faith that hereditary variation results from stochastic copying errors and unavoidable damage to the genome. In this perspective, the genome is a “read-only memory” (ROM) data storage system passed on from one generation to the next.<sup>1</sup>

In the past 60 years, since the structure of DNA was elucidated [5,6], molecular biologists have studied the basic mechanisms of long-term genome change. They have discovered a wide array of proofreading and damage repair biochemical systems that remove copying errors and correct DNA damage. At the same time, they have revealed an amazing and wholly unanticipated array of cellular and molecular systems that operate to generate genome variability, both temporary and structural. As we begin the second decade of the 21st century, accumulating empirical evidence has thus shifted the perspective on genome variation to that of an active inscription process changing the information passed on to future generations.<sup>2</sup>

In other words, we now have to reconsider the genome as a “read–write” (RW) information storage system highly sensitive to biological inputs. Although there are other cell structures that contain hereditary information [8,9], we know most about how susceptible genomic DNA is to biologically regulated modifications [10]. Indeed, I will argue that one of the main adaptive features of DNA-based heredity is that DNA is a highly malleable storage medium, permitting rapid and major changes to complex organisms without disrupting their functional integrity. Every time we do a molecular genetic intervention to work out the operation of some intricate cell control circuit, we make use of this malleability.

This review will summarize some of the molecular biology lessons acquired since 1953 about (1) genome structure, functions and organization, (2) the time scales for genome inscriptions, (3) proofreading and DNA damage repair, (4) cell action in restructuring genomic DNA (*i.e.*, natural genetic engineering or NGE), (5) the historical role of NGE processes as evidenced by the evolutionary DNA record, and (6) the conceptual and experimental challenges posed by the success of living cells in creating adaptive genomic novelties in the course of evolution.

As we shall see, the molecular lexicon and cellular writing methods are quite ample, at the levels of both temporary and permanent inscriptions on and in DNA. The references to the material in the tables will be available online in links at <http://shapiro.bsd.uchicago.edu/PLREV.RWGenome.html>. This format makes for ease of reading, keeps the review

<sup>1</sup> Sydney Brenner captured this view well from an informatics perspective in his Alan Turing Centennial tribute: “Turing’s ideas were carried further in the 1940s by mathematician and engineer John von Neumann, who conceived of a ‘constructor’ machine capable of assembling another according to a description. A universal constructor with its own description would build a machine like itself. To complete the task, the universal constructor needs to copy its description and insert the copy into the offspring machine. Von Neumann noted that if the copying machine made errors, these ‘mutations’ would provide inheritable changes in the progeny.” [4].

<sup>2</sup> Barbara McClintock most vividly expressed this view in her 1983 Nobel Prize lecture when she described her work analyzing the mutagenic effects of X-rays in maize: “The conclusion seems inescapable that cells are able to sense the presence in their nuclei of ruptured ends of chromosomes and then to activate a mechanism that will bring together and then unite these ends, one with another. . . . The ability of a cell to sense these broken ends, to direct them toward each other, and then to unite them so that the union of the two DNA strands is correctly oriented, is a particularly revealing example of the sensitivity of cells to all that is going on within them. . . . There must be numerous homeostatic adjustments required of cells. The sensing devices and the signals that initiate these adjustments are beyond our present ability to fathom. A goal for the future would be to determine the extent of knowledge the cell has of itself and how it utilizes this knowledge in a “thoughtful” manner when challenged. . . . In the future, attention undoubtedly will be centered on the genome, with greater appreciation of its significance as a highly sensitive organ of the cell that monitors genomic activities and corrects common errors, senses unusual and unexpected events, and responds to them, often by restructuring the genome.” [7].

within length constraints, and highlights to the interested reader the scope of empirical information underpinning the RW genome concept.

## 1. The genome is a formatted data storage system for living cells

### 1.1. *The complexity of cell reproduction and the magnitude of the data storage and data access challenges*

It is essential for scientists to keep in mind the astonishing reliability and complexity of living cells. Even the smallest cells contain millions of different molecules combined into an integrated set of densely packed and continuously changing macromolecular structures. Depending upon the energy source and other circumstances, these indescribably complex entities can reproduce themselves with great reliability at times as short as 10–20 minutes. Each reproductive cell cycle involves literally hundreds of millions of biochemical and biomechanical events. We must recognize that cells possess a cybernetic capacity beyond our ability to imitate. Therefore, it should not surprise us when we discover extremely dense and interconnected control architectures at all levels. Simplifying assumptions about cell informatics can be more misleading than helpful in understanding the basic principles of biological function.

Two dangerous oversimplifications have been (i) to consider the genome as a mere physical carrier of hypothetical units called “genes” that determine particular cell or organismal traits, and (ii) to think of the genome as a digitally encoded Read-Only Turing tape that feeds instructions to the rest of the cell about individual characters [4].

As we are learning from the ENCODE project data, the vast majority of genomic DNA, including many so-called “non-coding” (nc) segments, participate in biologically specific molecular interactions [11]. Moreover, the term “gene” is a theoretical construct whose functional properties and physical structure have never been possible to define rigorously. It is telling that genome sequence annotators used to call protein-coding regions (chiefly in prokaryotic DNA) “genes,” but now use the more neutral terms CDS, for “coding sequence.”<sup>3</sup>

The Turing tape idea falls short, as we will see, because it does not take into account direct physical participation of the genome in reproductive and regulatory interactions. The concept of a Read-Only Turing genome also fails to recognize the essential Write capability of a universal Turing machine [12], which fits remarkably well with the ability of cells to make temporary or permanent inscriptions in DNA.

### 1.2. *The specific functions a cell data storage system has to fulfill*

The genome is a particularly complex data and information storage system because it is an integral component of a living, reproducing cell. The special status of the genome has certain similarities with both hard or flash drives and random-access memories (RAM) in electronic information systems. However, the need to reproduce and change when necessary imposes additional requirements. We can distinguish at least seven genomic functions living cells must execute:

- (1) Compacting and packaging the DNA appropriately within the cell;
- (2) Replicating the genome in synchrony with the cell division cycle;
- (3) Transmitting replicated DNA accurately to progeny cells at cell division;
- (4) Maintaining the data files and formatting signals necessary for replication, transmission and accurate production of encoded RNA and protein molecules;
- (5) Appropriately regulating the timing and extent of coding sequence expression – often coordinating multiple dispersed locations, which encode protein and RNA molecules that function together;
- (6) Detecting and repairing replication errors and damage to the genome;
- (7) Restructuring the genome when necessary (either as part of the normal life cycle or in response to DNA damage or to a severe ecological challenge).

<sup>3</sup> I will use the terms “genetic locus” and “genetic loci” here in the way that my colleagues tend to use the term “gene.” I prefer genetic locus because it is an operational definition, determined by *in vivo* genetic mapping experiments or DNA sequence analysis. We know that processes like differential transcription initiation or termination and differential splicing confer on each genetic locus the capacity to encode related but distinct products, which may have quite different functions. Thus, inclusion of the term locus ties the terminology to specific regions of the genome without any functional presuppositions.

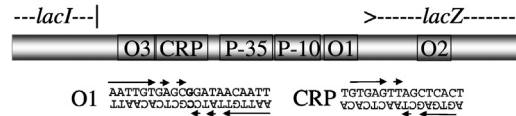


Fig. 1. *lac* operon transcriptional control region. The signals include a bipartite (P-35, P-10) *lacP* promoter for interaction with RNA polymerase and initiation of transcription into messenger RNA (mRNA), a *CRP* binding site for the cAMP–CRP complex that stabilizes the RNAPol–*lacP* complex in the presence of the cAMP second messenger (indicating glucose is not available), a transcription terminator downstream of the operon (not shown), a ribosome binding site upstream of *lacZ* (not shown) that facilitates translation of the mRNA into protein, ATG start and TGA or TAA stop codons for each CDS, and 2 of the 3 *lacO* operator sites for coordinated binding of two LacI-encoded repressor dimers to form a DNA loop blocking access to *lacP* when lactose is not available.

Functions (1)–(6) and sometimes (7) are all necessary for survival, completion of the normal life cycle and successful reproduction. Function (7) is essential for changing to survive major ecological upheavals, when the previously encoded genomic data prove insufficient for growth and reproduction.

### 1.3. DNA is a highly formatted data storage medium and information expression system

We are familiar with the idea of genomes containing coding sequences for RNA and protein molecules. We can designate this the CDS database, digitally encoded in nucleotide sequences. However, the ability of the cell to access CDS data is completely dependent upon additional digital formatting of the coding sequences with signals for interactions with other cell molecules and signal transduction networks [13].

To take a relatively simple, classic example, proper expression of the *E. coli lac* operon encoding proteins needed for lactose utilization [14,15], the properly regulated expression of the contiguous *lacZ* and *lacY* coding sequences for beta-galactosidase and beta-galactoside permease depends upon formatting by a number of DNA signals (see Fig. 1).

This simple prokaryotic example is far less elaborate than eukaryotic genetic loci, where coding sequences are often split into **exons** (expressed CDS sequences) and **introns** (intervening sequences). These require proper signals at each exon/intron junction for splicing the discontinuous CDS regions in the initially transcribed RNA into a single continuous CDS in the resulting spliced mRNA (Fig. 2, see Section 1.8 below).

Further complexities arise when control of transcription by cell circuits requires the interaction of multiple “transcription factor” (TF) proteins complexed with enhancer/inhibitor regions that may be at some distance from the transcribed region. These “*cis*-regulatory modules” (CRMs) generally contain sequences for the cooperative binding of multiple TF molecules [13,16].

The elaborate formatting of transcription and its control permit cells to execute CDS expression in a flexible and highly controlled manner, both in response to variable outside inputs as well as during the execution of developmental cell differentiation programs. A number of authors have emphasized the algorithmic and computational nature of the regulatory regimes involving CRMs [17]. They permit the execution of elaborate multicellular morphogenetic routines as well as flexible responses to countless unpredictable contingencies.

### 1.4. Genome composition

Genomes are composed of diverse classes of DNA sequence elements. There are not one but many genetic codes affecting all the seven genomic functions listed above. Deciphering all these codes and understanding how they contribute to genome functions is the goal of the ENCODE project (*Encyclopedia of DNA Elements*) [18]. The decision to call all these various genome components simply “DNA elements” reflects the wisdom gained from attempts to deconstruct and define “genes” and other presumptive basic genome units.

Many DNA elements are repeated in the genome [19,20]. Some are localized with many copies together (often called “satellites” or “tandem repeats”), while others are dispersed at different sites across the genome. Table II-2 of [10] presents a summary of some different classes of repetitive DNA (<http://shapiro.bsd.uchicago.edu/TableII.2.shtml>).

The existence of so many different kinds of repeat element lead us to two important conclusions about the evolution of genome architecture:

*1.4.1.* Repeats arose from active amplification processes. In the human genome, for example, the initial draft genome revealed 44% dispersed repeats (over 3 million copies) [21], and more recent analysis pushes that fraction close to 67% [22]. The protein-coding sequences (exons) amount to less than 2% of the human genome. We know the molecular processes by which these dispersed repeats have been copied and distributed through the genome. Thus, the vast majority of the human genome and significant fractions of other genomes (<http://shapiro.bsd.uchicago.edu/TableII.1.shtml>) have been inscribed by well-defined molecular processes.

*1.4.2.* Much of the genome is composed of defined DNA elements that actually constitute complex **data cassettes**, comprising coding sequences, transcription signals, splice sites, and other classes of functionally significant information [23]. Regardless of where they insert in the genome, these cassettes will have significant and predictable effects on the functioning of nearby DNA sequences [24]. In this regard, the evolution of genome architecture can be compared to a Lego-like or a plug-in motherboard construction process.

Shortly after repetitive DNA was discovered as an abundant genome component, Britten and Davidson proposed that repetitive elements could constitute distributed regulatory sites and so form intra-genomic networks uniting distant genetic loci [25].

### *1.5. Functional genome formatting for DNA compaction and localization in the nucleus*

We understand the most basic level of genome compaction best in eukaryotic cells. This is the winding of  $\sim 147$  base-pairs of DNA around a core of 8 positively charged “histone” proteins into a structure called a “nucleosome” (nuclear body) [26] (Fig. 3, see below). Genomicists have deciphered a DNA sequence code which helps determine where nucleosomes form, but their positions are subject to active modification under particular conditions [27,28]. The nucleosomes can form at different densities along the DNA and then fold into higher-order “chromatin” complexes, which may involve additional proteins and RNA molecules [29].

The name “chromatin” derives from cytogenetics, where chemical stains were used to visualize chromosomes in the cell nucleus. Lightly staining “euchromatin” regions have a lower density of nucleosomes and are considered to contain active DNA. More intensely staining “heterochromatin” regions have a higher density of nucleosomes and were considered to contain DNA inactive for replication and transcription. Regions of active euchromatin, silent heterochromatin and other distinct types of chromatin can be separated from each other by DNA elements known as “boundary elements” or “insulators” [30]. Insulator elements also have the ability to disrupt the functional connections between CRMs and promoters in transcriptional control. Some insulators exert their effects by binding to the surrounding nuclear lamina [31]. Such attachment appears to separate DNA segments into distinct functional domains.

As we are learning, thanks to new microscopic and labeling technologies, the eukaryotic nucleus is a highly organized and sub-compartmentalized organelle [32]. There are special sub-nuclear foci, sometimes called “factories,” where discrete genome functions take place, such as replication, transcription, and repair [33,34]. This means that the dynamic spatial organization of the genome has functional significance, and it should be no surprise that the genome is formatted accordingly. For example, there are other signals called S/MARs (surface/matrix attachment regions) that are also sites for localization to the nuclear lamina [35]. It is important to note that LINE elements, one class of dispersed repetitive DNA that comprises about 20% of the human genome [21], have twice the average density of S/MAR sequences [36]. Thus, these and other repetitive elements play a major role in the functional architecture of the genome.

### *1.6. Formatting for replication for transmission to daughter cells*

Genomic DNA occurs in different configurations in distinct cell types [37]. Moreover, genome sizes range over more than five orders of magnitude ( $\sim 0.3$ –100,000 MB) [38]. Such differences in genome configuration dictate a variety of formatting requirements for replication. It is necessary to specify the origins where replication starts, the structures where replication ends on circular DNA, make sure that both strands are fully replicated, and provide special structures for the ends of linear molecules to ensure their stability (Table 1).

Once genomic DNA molecules have been replicated (and the two copies separated), the genomes must be moved to each of the two daughter cells at division. This movement is accomplished by attachment to microtubules in eukaryotes or to tubulin- and actin-like molecules in bacteria. In eukaryotes, microtubule attachment occurs at complex

Table 1

Genome formatting for replication (available with references at <http://shapiro.bsd.uchicago.edu/Table1F.GenomeFormattingforReplication.html>).

Formatting signal	Function
Origin of replication ( <i>Ori</i> )	Permit formation of initiating replication complex on duplex DNA
Terminus ( <i>Ter</i> )	Allow completion of circular DNA replication
Site-specific recombination site ( <i>dix, par</i> )	Permit the resolution of dimeric circular replication products into two monomeric circles for partition to daughter cells
Okazaki primer initiation sites	Facilitate RNA primer synthesis on lagging strand
Telomeric hairpin loop with resolution sites	Continuous DNA strand at the end of a linear genomic molecule (prokaryotes) to permit complete replication of the lagging strand end and separation into two linear copies
Telomere	Specialized structure at the end of eukaryotic linear molecules that allow extension of the lagging strand by the telomerase reverse transcriptase

structures called “kinetochores” [39]. The kinetochore forms at a special chromosome region called the “centromere” (central body). The centromeres of most eukaryotic chromosomes have particular genetic and chromatin configurations. Typically, centromeres are composed of one or more families of tandem repetitive element, and centromeric nucleosomes contain special histone variants [40].

### 1.7. Formatting for transcription initiation, elongation and termination

We have already mentioned recognition sites for RNA polymerase (promoters) and transcription factors in generalizing on the *lac* operon example of transcriptional formatting. Chromatin formatting and the three-dimensional organization of the genome also have a profound impact on transcriptional activity [41].

An added dimension of formatting promoters and *cis*-regulatory modules (CRMs) is the existence of coordinately regulated suites or networks of multiple genetic loci distributed throughout the genome (sometimes called “regulons” or “*cis*-regulatory networks”) [42]. These dispersed but functionally connected loci typically encode multiple protein and RNA products that work together in a particular cellular task. The CRMs serve to connect individual genetic loci to the same signal transduction networks, which activate or inhibit specific transcription factors. Shared CRMs thus integrate CDS expression control into cell signal response networks. The presence of cognate CRMs at distant loci raises the question of how they came to be distributed in evolution [43]. Before we knew about the mobile genetic elements described below, it was virtually impossible to answer this question. But now we can, and we will return to this question later.

In addition to formatting that helps regulate the initiation of transcription, there are also sequence determinants that format transcript elongation and termination. The regulation of transcript elongation has been far less studied than initiation, but we know it can be controlled because the LINE-1 dispersed repeat has a profound retarding effect when it is present [44]. Since transcript elongation must pass through introns, we cannot ignore the effect of these so-called “non-coding” sequences on genome expression.

### 1.8. Formatting for post-transcriptional RNA processing and editing

Once transcribed, RNA molecules undergo a number of subsequent modifications that are formatted in the genome. We will see later that the addition of a generic A-rich sequence to the 3' ends of many different kinds of cellular RNAs plays a key role in facilitating the movement of sequences from one genomic context to another.

We have discussed splicing in the context of invalidating the “gene” concept. There are different kinds of splicing processes, but here we will only consider the best-characterized one, carried out by the “spliceosome” ribonucleo-protein structure [45]. This form of splicing requires a variety of strictly- and loosely-defined formatting signals that involve the ends of the two exons as well as various sites in the intron (Fig. 2).

“Trans-splicing” occurs by joining exons from two transcripts to produce chimeric mRNAs, some encoding chimeric proteins [46]. Trans-splicing is a regular feature of the genome expression process in certain organisms [47], and it serves as a mechanism for protein diversification by creating novel exon combinations in many taxa [48]. Recent data indicate that trans-splicing occurs more frequently than previously believed [49].

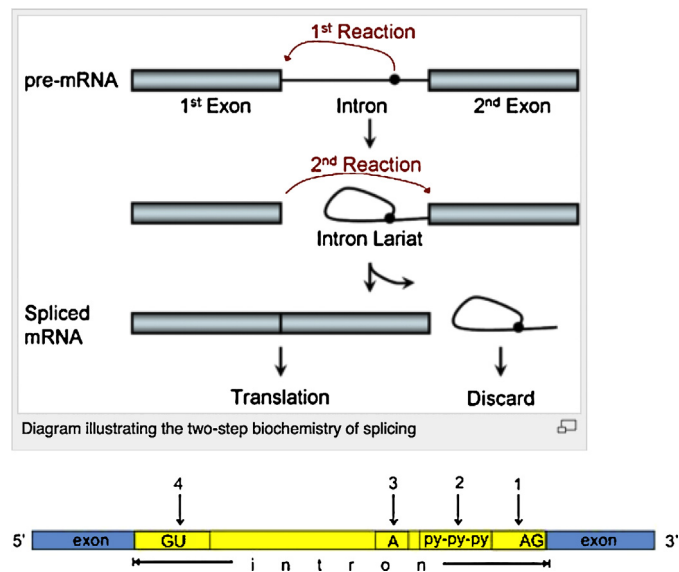


Fig. 2. Spliceosomal action and formatting. (#1) Splice acceptor ( $S_A$ ) signals at the 3'-end of each intron and the first nucleotide of the downstream exon, almost always N-C-A-G-{cut}-G. (#2) A "polypyrimidine" tract 20–50 nucleotides upstream of the  $S_A$  AG. (#3) A "branch point" C-U-R-{A}-Y upstream of the polypyrimidine tract; this is where the 5'-end of the intron joins in an unusual 5'-2' phosphodiester bond to the {A} to form the excised intron "lariat" structure. (#4) Splice donor ( $S_D$ ) signals at the 5'-end of each intron, almost always GU; the  $S_D$  includes sequence elements from the upstream exon and the intron (M-A-G-{cut}-G-U-R-A-G-U). Reproduced with Creative Commons rights from Wikipedia.

In addition to splicing, there are two more kinds of post-transcriptional processing of RNA molecules. One is cleavage into smaller RNAs by endonucleases, and this cleavage can be either sequence- or structure-specific [50]. Another kind of post-transcriptional processing is called RNA editing [51]. RNA editing mostly results from deamination of C to U or A to I (inosine), both of which are coding transitions and thus modify the information content of the RNA molecules. There are also cases where nucleotides are added to or deleted from the transcripts. Both deamination and nucleotide addition are enzymatic processes subject to sequence specificity mediated by enzyme binding to RNA substrates. RNA editing has been adapted, among other functions, as a regulatory mechanism to modulate ion channels in mammalian neurons [52].

The significance of RNA editing and other post-transcriptional RNA modifications is that existing genome sequences can be altered at the transcript level and then undergo reverse-transcription and reinsertion into the genome [53]. Thus cells have a means of generating *ex post facto* genome inscriptions that occur following expression and possible influence on phenotype. In particular, trans-splicing and reverse transcription provide a realistic molecular mechanism for combining sequences from distant regions of the genome into novel chimeras (see Section 5.2).

## 2. Different types of genome inscriptions occur over three distinct biological time scales

For the purposes of my argument in favor of the RW genome, it helps to recognize at least three distinct biological time frames for genome function and for the different modes of genome inscriptions at each of them [10,36].

### 2.1. Within the cell cycle

This shortest biological time-scale includes all the operations needed to proceed from one cell division to the next. There are constantly changing nucleoprotein complexes involved in transcribing RNAs from genomic DNA, in adjusting the transcription apparatus to respond to environmental and metabolic inputs, as well as to maintain progress through the cell cycle, replicating the genome, preparing the genome for equal distribution at cell division, and completing cell growth and separation [13,16,54]. These nucleoprotein complexes constitute short-term inscriptions on the genome, each providing the cell with instantaneous information about the state of one or more genomic regions. Table 2 lists some of the processes we understand chiefly in terms of these short-term nucleoprotein inscriptions.

Table 2

Processes governed chiefly by transient nucleoprotein complexes (available with references at <http://shapiro.bsd.uchicago.edu/Table2A.ProcessesGovernedByNucleoproteinComplexesWithintheCellCycle.html>).

---

Metabolism and biosynthesis ( <i>e.g.</i> , adjustment to nutritional & environmental changes)
Response to intercellular signals, such as pheromones and hormones
Damage repair
Programmed cell death (apoptosis) decisions in multicellular development and damage response

---

Cell-cycle control is the best illustration of the transient effects of nucleoprotein complexes [55]. Progress through the cell cycle is monitored by a continuous evaluation of genome nucleoprotein complexes. If imbalances or abnormalities occur, surveillance systems activate a series of “checkpoint” controls to make sure that the cell cycle completes properly [56]. Once a checkpoint detects a problem, it communicates with the cell cycle control circuits to halt progress until the problem has been resolved [57].

There are checkpoints for numerous aspects of cell reproduction: nutrition [58], DNA replication [59,60], telomere capping [61], DNA damage [62], cell growth and size [63], chromosome alignment on the mitotic spindle [64], and spindle orientation [65]. Clearly, keeping all the different processes needed for the cell cycle in proper order and ensuring their completion without damage involves constant short-term reading and writing on the genome.

## 2.2. Over multiple cell cycles

Many biological processes occur over multiple cell cycles. There are basically two life histories that involve a large but limited number of cell generations: (i) the development of a multicellular organism from an individual progenitor, typically a fertilized egg, and (ii) the vegetative growth of a population of related cells in a particular ecological niche. Two distinct forms of genome inscription characterize these different situations, and they are not mutually exclusive.

### 2.2.1. Inscription over multiple cell cycles in multicellular development by epigenetic modifications

The principle mode of multi-cell cycle genome inscription in the development of multicellular organisms is labeled “**epigenetic**.” Conrad Waddington articulated the modern concept of epigenetic inheritance before we had any clear idea of the molecular basis of how heritable traits were transmitted from parent to progeny [66]. He argued that there had to be a heritable overlay (*i.e.*, an “epi” layer) on the normal trans-generational “genetic” process to explain the maintenance of cell differences in the development of tissues and organs in multicellular organisms, where quite distinct phenotypes were displayed by cells carrying the same genomes.

It is a fascinating feature of the history of molecular cell biology that our current understanding of the epigenetic role of chromatin formatting and modification has come to fit Waddington’s prescriptions so well. We now envision epigenetic formatting of the genome as occurring by a sequence of physico-chemical operations: (i) DNA methylations, (ii) compaction of the DNA chains by tightly winding them around positively charged proteins called “histones” (H1, H2A&B, H3, H4) into structures called “nucleosomes” (147–148 base-pairs wound 1.75 times around the histone octamer [67]) and (iii) chemical modifications of specific amino acid residues on the exposed “tails” of the histone proteins of the nucleosome structure (Fig. 3).

The different chemical modifications on their exposed tails constitutes a “histone code” [69] that can be read by control circuits in the cell. These modifications are dynamic and often reflect a particular process in progress or just completed. Regions with ongoing transcription acquire specific histone modifications [70]. In addition to effects on transcription, histone modifications serve as signals for the regulation of other genomic processes. For example, H2BK123 monoubiquitylation is needed for DNA repair [71], H3K36 methylation enhances NHEJ repair [72], H3K56 acetylation signals the end of DNA repair [73], symmetrical dimethylations of H2AR3 and H4R3 characterize human satellite repeats [74], and H3K9 methylation aids centromere cohesion [75].

There are also histone protein variants for special genomic states. For example, H2AX occurs in phosphorylated form at sites of double-strand DNA breakage [76], H2AZ characterizes the genomes of pluripotent cells [77–79], and centromeres contain a specialized histone called CENH3 [80,81].

The transition from one multicellular generation to another by gametogenesis and fertilization is now seen as a major period of epigenetic reprogramming (*i.e.*, inscription and erasure) [82]. A key indication of this epigenetic



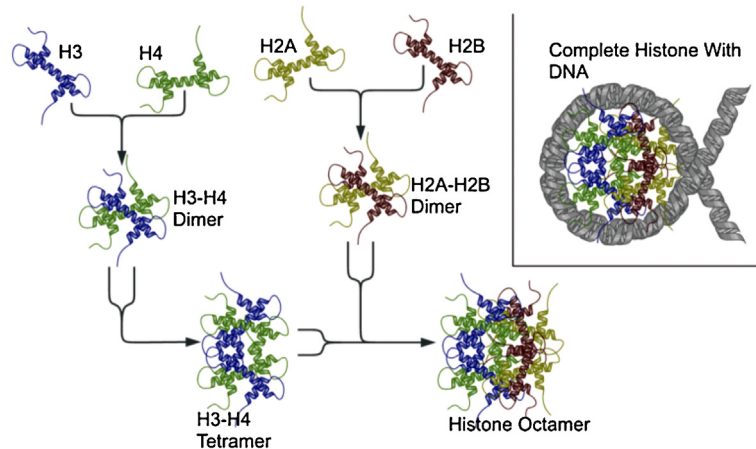


Fig. 3. Schematic diagram of nucleosome assembly and structure. Note the tails on the individual histones emerging from the complexes. Specific amino acids in these tails are the targets of chemical modification to execute the histone code [68,69]. By Richard Wheeler (Zephyris), reproduced with Creative Commons rights from Wikipedia.

reprogramming is the phenomenon of “imprinting,” by which the expression of a genomic region in the somatic tissues of the next generation depends on the parent who contributes it to the zygote [83].

During multicellular development, pluripotency is lost, stem cell identities are established, and terminally differentiated states achieved by successive waves of changes in chromatin organization [84]. In one of the most exciting areas of biotechnological application, our ability to derive pluripotent stem cells from differentiated cells demonstrates that epigenetic inscriptions can be rewritten with the appropriate stimuli [85].

RNA molecules target epigenetic formatting to sites in the genome with complementary sequences [86]. One of the basic modes of small RNA-mediated regulation occurs in the form of transcriptional silencing by so-called RITS complexes (RNA Interference by Transcriptional Silencing). The heart of an RITS complex is a member of the “argonaute” protein family, which binds the small RNA (on the order of 2 dozen nucleotides) and helps it find the appropriate genome homologies, probably by forming RNA:RNA duplexes with nascent transcripts [87]. The RITS complex then recruits chromatin-formatting functions to execute the appropriate DNA methylation and histone modifications.

Generally, RNA-directed chromatin formatting has been studied in the context of genome silencing, but we still have a great deal to learn about this system. The animal germ-line contains a special subfamily of argonaute-like activities called “Piwi” proteins, after the first such molecule discovered [88]. These germ-line-specific proteins bind so-called “piRNAs,” and the resulting RITS complexes are essential for germ-line development, gametogenesis and the inhibition of mobile element dispersed repeats in the genome [89]. This silencing effect of piRNA-directed chromatin is one of the key guardians of germ-line genome stability [90]. Recent discoveries show that eukaryotic cells acquire fragments of invading viral and mobile element DNAs and place them in special loci dedicated to piRNA formation [91]. An equivalent genome protection system by DNA acquisition from invaders occurs in the bacterial and archaeal CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) systems. Experimental work in these prokaryotic systems has shown that DNA acquisition from infecting viruses is rapid enough to generate infection-induced resistance [92]. This is an important form of rapid genome inscription that provides a form of acquired immunity against viruses, mobile elements, and other invaders. It invalidates the universality ascribed to the Luria–Delbrück experiment [93] as demonstrating that virus infection cannot induce resistance.

### 2.2.2. Over multiple cell cycles in vegetative populations by structural DNA changes

In addition to epigenetics, direct restructuring of genomic DNA serves as a reversible form of inscription over multiple cell generations in vegetative cell populations. There are diverse examples and mechanisms for such intermediate-term DNA rewriting. Some bacteria use various reversible DNA modifications to turn expression of certain functions, typically surface proteins, on and off. This process has come to receive the name “phase variation” [94].

Using a variety of molecular mechanisms, both bacteria and eukaryotic microbes can alter the DNA coding sequences for surface proteins. Since this process is generally used for evading recognition and destruction by the

adaptive immune system, it has received the name “antigenic variation” [94]. A table of both phase and antigenic variation by DNA rearrangements is available online at <http://shapiro.bsd.uchicago.edu/TableII.5.shtml>.

An illuminating example of multi-cell cycle DNA changes that do not alter the germ-line genome occurs in the ciliated protozoa, or ciliates. The ciliates comprise a large family of eukaryotic microbes that carry both germ-line and somatic nuclei in the same cell [95,96]. During vegetative growth by mitotic cell division, the larger somatic “macronucleus” encodes the RNA and protein molecules needed for reproduction, and the small germ-line “micronucleus” maintains the complete genome needed for sexual reproduction. Macronuclear somatic chromosomes are small (often carrying a single coding sequence) and in high copy numbers. Micronuclear germ-line chromosomes are typical eukaryotic chromosomes and diploid. Micronuclear germ-line chromosomes carry many sequences absent from the somatic macronucleus.

When nutrition is limited, ciliates undergo meiosis and mating. Afterwards, the old macronuclei degenerate, and a new macronucleus develops from a descendant of the post-mating zygote nucleus. Forming a new macronucleus is guided by RNA molecules from the old macronucleus [97]. The process is complex and involves elimination of much germ-line-specific DNA (up to 90%, depending on the species), as well as reassembly of scrambled germ-line sequences into accurate coding sequences [95,98]. The thousands of DNA cutting and splicing events involved in development of each new macronucleus occur regularly and reliably in the space of a few hours. The DNA acrobatics of macronuclear development constitute the most elaborate example known of rapid restructuring to form a novel functional genome architecture within a single cell generation.

### 2.2.3. DNA changes in multicellular development: somatic nervous system diversification by activation of mobile genetic elements

There are two cases where activation of mobile genetic elements has been shown to occur in animal nervous systems: LINE retrotransposons in mice [99,100] and multiple classes of mobile element in *Drosophila* [101]. The evidence suggests that somatic transposition and retrotransposition events individually modify neurons to add diversity to the node structure of the resulting neural networks.

## 2.3. Over evolutionary time

Evolutionary inscriptions are those written into the DNA sequence structure that influence taxonomic divergence. Although the traditional concept of gradual genome change in evolution assumes that these inscriptions have to accumulate within vertical lineages over long chronological periods, some evolutionary inscriptions clearly occur in a single organismal generation by cell actions outside normal vertical heredity. Here, we will discuss those inscriptions that involve interactions between at least two different cells. In Section 4, we will consider the intracellular natural genetic engineering processes that restructure genomes to facilitate evolutionary innovation.

### 2.3.1. Symbiogenesis

Symbiogenesis is the formation of new cell types or organisms with novel genome mixtures by cell fusions [102–104]. Based on ribosomal RNA taxonomy [105], we know today that symbiogenesis was responsible for two of the most important events in all of evolutionary history:

- (a) Incorporation of an *alpha-proteobacterium* as the mitochondrion in the ancestor of all eukaryotic cells [106].
- (b) Incorporation of a *cyanobacterium* as the plastid in the ancestor of photosynthetic eukaryotes [107].

Secondary and tertiary endosymbiosis involving both red and green algae are further responsible for the evolution of a number of additional photosynthetic eukaryotic lineages [107–109]. Secondary endosymbiogenetic organisms contain at least four separate genomic compartments: nucleus, mitochondrion, plastid and nucleomorph (the former nucleus of the endosymbiotic alga).

After a symbiogenetic fusion, there is extensive intracellular transfer of DNA from the organelle compartments to the nuclear genome [110,111]. Thus, most of the proteins needed for plastid and mitochondrial function are synthesized from nuclear transcripts in the cytoplasm and have to be reimported into the organelle where they act [112].

Table 3

Documented Horizontal Transfers in the three domains of life (available online with references at [http://shapiro.bsd.uchicago.edu/Table2C\(2\).DocumentedHorizontalDNATransferEventsintheThreeDomainsofLife.html](http://shapiro.bsd.uchicago.edu/Table2C(2).DocumentedHorizontalDNATransferEventsintheThreeDomainsofLife.html)).

---

Among prokaryotes (bacteria and archaea)
Among eukaryotes
Eukaryotes to prokaryotes
Microbes to multicellular organisms

---

### 2.3.2. Horizontal DNA transfer

Horizontal DNA transfer occurs when DNA passes between the cells of different organisms [113]. This form of lateral inheritance was first recognized as a major factor in the rapid evolution of bacterial antibiotic resistance [114]. Genetic, genomic and metagenomic studies have documented that horizontal DNA transfer is fundamental to evolution and involves widespread transfers between all domains of life (Table 3 and <http://shapiro.bsd.uchicago.edu/TableIII.1.shtml>) [10].

### 2.3.3. Hybridization

The term “hybridization” applies to matings between individuals from different populations or species. Hybridization is a powerful force for stimulating rapid genome-wide changes [115–118]. Since hybridization is the fusion of two haploid cells (gametes) from distinct organisms that usually do not interbreed, it has some analogies with symbiogenesis. The result is a cell with two haploid genomes that differ in their epigenetic configurations [119]. Combined epigenetic and genomic instability is a typical result of hybridization [120].

As an interesting commentary on the documented sources of taxonomic divergences, it should be noted that artificial selection has never produced a novel species. In contrast, new species have long been produced by interspecific hybridization in agriculture [116,117]. Moreover, the genome sequence record provides abundant evidence for hybridization as a powerful driver of “hybrid speciation” in plants and animals [121,122]. Darwin’s “abominable mystery” of rapid seed plant (angiosperm) diversification results from the ease of interspecific hybridization and the capacity for self-fertilization by hybrids in these organisms [123].

Hybridization has several distinct effects on genome composition:

**2.3.3.1. Phenotypic diversification** Hybrids often display novel characters that are not just the average of the two parents [124]. In the wild, this provides them with potential selective advantages [125]. In domestication, the novel characters have proven useful to the hybridizers. The sources of these novel characters are often unknown, but changes in transcriptional regulation, epigenetic regulation and genome structure have all been documented in hybrid progeny [126].

**2.3.3.2. Changes in chromosome composition** Hybridization changes the karyotype (nuclear composition) in a variety of ways. Changes in total chromosome numbers are called polyploidy (more than two complete haploid genomes per nucleus) [127] and aneuploidy (an abnormal number of chromosomes that is not an integral multiple of the haploid genome). The most common change in hybrids is whole genome duplication (WGD), also called tetraploidy [128]. WGD ensures successful meiosis and the capacity to produce gametes in hybrids. WGD can happen at very high frequencies (25% of all hybrids in a recent *Arabidopsis* study [129]). In our own ancestry, there were two successive WGD events at the beginning of vertebrate evolution [130–132].

**2.3.3.3. Genome variability** The genomes of hybrids display great instability. The repetitive components of the genome are particularly subject to instabilities, probably because the epigenetic controls that stabilize them have been disrupted [133]. But ongoing genome instability also results in chromosome rearrangements and chromosome losses that reduce the chromosome complement back to the former diploid number [127].

This “diploidization” process leaves a residue of so-called “segmental duplications” distributed and rearranged in new locations throughout the genome [134]. In fact, the conclusion that a sequenced genome has resulted from a WGD event is generally based on the observation of numerous segmental duplications from many ancestral locations in the prior diploid configuration.

The fact that segmental duplications are separated, frequently on different chromosomes or inverted in orientation on the same chromosome [135], tells us that processes of chromosome breakage and rejoining have been an integral part of diploidization. Among the most variable components of the hybrid genome are repetitive DNA elements [136]. Since repetitive DNA plays a fundamental role in genome formatting [36], these changes contribute to the establishment of a new “genome system architecture” [137,138] in the diploidized hybrid progeny.

**2.3.3.4. Protein family and network amplification** An important consequence of WGD is that all genome components are duplicated and become available for change without losing the original function, as emphasized by Ohno [139]. WGD contributes to the amplification and elaboration of protein-encoding loci. But WGD also amplifies higher-order components, such as the Hox and ParaHox complexes [140,141].

The freedom for change in duplicated copies following WGD also applies to cell networks encoded by sets of dispersed genetic loci. The reason is that distributed genomic networks can undergo rapid modification and come to adopt encode novel functions without loss of the original functionality. This potential has contributed to the adaptation of signal transduction and other multicomponent systems to diverse cell and developmental functionalities [142–145].

#### 2.3.4. The roles of viruses and the virosphere

Numerically, viruses are the most abundant biological entities on the planet [146]. Scientists conceiving the genome as ROM have typically considered viruses to be merely parasites. But the RW perspective encompasses viruses as active agents of variability, as intercellular vectors for nucleic acid transfer, and as laboratories for evolutionary experimentation (<http://shapiro.bsd.uchicago.edu/ExtraRefs.Virosphere.shtml>) [147].

Metagenomic studies of environmental samples show that viruses contain numerous classes of cellular sequences unrelated to viral reproduction [148]. Virtually all cells contain endogenous proviruses and viral fragments in their genomes, where they exert important effects on phenotype and genome formatting [149]. In our own evolution, it is widely accepted that endogenous retroviruses played an essential role in placental evolution, a key step in the divergence between marsupials and eutherian mammals ([http://shapiro.bsd.uchicago.edu/Retroviral\\_involvement\\_in\\_placenta\\_evolution.html](http://shapiro.bsd.uchicago.edu/Retroviral_involvement_in_placenta_evolution.html)) [150].

The recent discovery of a class of giant viruses with mega-base-pair sized genomes has had a tremendous impact on thinking about the evolutionary role of viruses [151]. These mega-viruses infect amoeba, algae and an unknown array of other cell types [152]. Their genomes contain sequences from all three domains of life [153]. The presence of these viruses in amoeba and related eukaryotic microbes has led to the proposal that these cells constitute an evolutionary “melting pot” for the generation of novel chimeric sequences and DNA segments [154]. A recent report tells of the discovery of specialized subviral parasites (“virophages”) and virus-specific transposons (“viroposons”) in this fascinating amoebal multi-genomic ecology [155].

### 3. Cells and their genomes are not passive victims of replication errors or DNA damage

Since McClintock’s pioneering cytogenetic studies of chromosome break repair (footnote 2), molecular genetics has developed a whole field devoted to analyzing the proofreading and correction of replication errors as well as the cellular repair of DNA damage. The literature is vast. Here we will focus on aspects relevant to the reduction of accidental changes and the cell’s biochemical action in damage-induced mutagenesis.

#### 3.1. Replication proofreading

The first step in proofreading activity operates during replication and is denoted “exonuclease proofreading” [156]. In exonuclease proofreading, the DNA replication complex arrests when it has made an incorporation error, switches to an exonuclease function that removes the incorrect nucleotide from the end of the nascent strand, and then resumes polymerizing. Exonuclease proofreading removes 99.9%–99.99% of incorporation errors [157].

A second postreplication proofreading system follows the replication fork and detects any remaining errors. This is “mismatch repair” [158], which depends upon monitoring proteins that sense distortions of the double helix at improper base-pairs, bind to them, and recruit addition proteins to remove the incorrect new strand so that it can be replaced by repair polymerization [159]. Mismatch repair catches 99%–99.9% of the errors that escaped exonuclease proofreading.

Under optimal growth conditions, an *E. coli* cell replicates its DNA at about 2 kHz ( $\sim 2000$  base-pairs per second). The polymerase has an intrinsic error frequency of  $\sim 10^{-5}$ . With the two-step proofreading system, this is reduced to  $< 10^{-9}$ . Note that this amazing precision results from sensing functions and the recruitment of correction activities, much as McClintock described for chromosome breakage repair.

### 3.2. Responses to genome damage

Cell DNA is subject to radiation and chemical damage. Sometimes the damage cannot be repaired, and cell death results. In addition, radiation and chemical damage are mutagenic in the surviving population, and a great deal of research on the response to DNA damage has been devoted to tracking down the sources of the induced mutations. Although initial thinking was that damaged DNA is simply a poor template for accurate replication, we now realize that mutations result from cell repair processes, not as an intrinsic consequence of damage.

The clearest experiments came in the 1950s from the Swiss physicist-turned-molecular-biologist Jean Weigle (1901–1968). Weigle used the bacterial virus  $\lambda$  (lambda) as a test system to explore UV lethality and mutagenesis. By studying killing and mutation of  $\lambda$  virus, Weigle could separate the DNA being examined (inside the virus particles before infection) and the cell environment where killing, mutagenesis and repair all took place after infection. Weigle systematically looked at all possible combinations of untreated and UV-irradiated  $\lambda$  virus and host *E. coli* cells. He obtained the following results [160,161]:

- (a) Untreated virus infecting untreated cells  $\Rightarrow$  no killing, basal virus mutations
- (b) Irradiated virus infecting untreated cells  $\Rightarrow$  killing, elevated virus mutations
- (c) Irradiated virus infecting irradiated cells  $\Rightarrow$  less killing, even more virus mutations
- (d) Untreated virus infecting irradiated cells  $\Rightarrow$  no killing, elevated virus mutations

Experiment (a) was the control situation, and experiment (b) confirmed that UV has both lethal and mutagenic effects on the  $\lambda$  virus. These results were expected. The surprises (discoveries) came in the last two experiments. Experiment (c) demonstrated that radiating the cells induced two different activities. One activity removed (repaired) lethal damage to the irradiated virus DNA. This activity was called “Weigle reactivation” [162]. The second activity increased the yield of mutations from irradiated DNA. This activity was called “Weigle mutagenesis” [163]. Experiment (d) was the biggest surprise. Irradiated cells produced additional mutations on **untreated** viral DNA. This “untargeted mutagenesis” [164] represented a totally unexpected inducible cell capacity to alter DNA sequences undamaged by radiation.

There are both error-free, non-mutagenic repair processes and error-prone, mutator processes (Table 4). More recent experimental results confirm the conclusion from Weigle’s early experiments that mutator repair functions cause damage-induced mutations. In particular, genetic ablation of various “error-prone” repair functions eliminates specific aspects of the mutagenic response [165].

### 3.3. Apoptosis decisions in response to DNA damage

Another means of avoiding irreparable DNA damage is for the damaged cell to undergo programmed cell death (apoptosis). An apoptosis option is present in all types of cells [166,167]. If DNA damage cannot be corrected or bypassed, then cell death is inevitable. But most cells execute the apoptosis decision based on the influence of signaling molecules in the environment [168]. Certain cell–cell signaling molecules serve as “death factors” and promote apoptosis (e.g., tumor necrosis factor, TNF [169]), while other signals generally described as “growth factors” bias the decision towards cell cycle arrest and repair [170].

## 4. Cell action in genome changes: Natural genetic engineering

### 4.1. Basic biochemical functions for cutting, splicing and otherwise altering DNA

Cells possess the biochemical activities that allow them to restructure DNA in the same ways that we do in laboratory genetic engineering (Table 5). Thus, the summary term “natural genetic engineering” (NGE) is suitable for

Table 4

**Some mutagenic and non-mutagenic repair processes** (available online with references at <http://shapiro.bsd.uchicago.edu/Table3B.SomeMutagenicandNon-mutagenicRepairProcesses.html>).

DNA damage type or agent	Error-free repair	Error-prone/mutagenic repair
Ionizing radiation, double-strand (DS) breakage, replication fork collapse	Homologous recombination (Rec)	Non-homologous end-joining (NHEJ)
UV radiation (thymine dimers)	Nucleotide excision repair (NER)	Lesion bypass repair
Alkylation damage	Base excision repair (BER), NER, de-alkylation	Lesion bypass repair
Large chemical adducts		Lesion bypass repair (DNA PolIII, PolIV & PolV in <i>E. coli</i> )
Oxidative damage	BER	Lesion bypass repair
Cytosine deamination to uracil	BER (uracil-N-glycosylase)	BER (uracil-N-glycosylase) – note that BER removal of uracil from DNA can have both mutation-prevention and mutation-promotion consequences

Table 5

Cell biochemical activities used in Natural Genetic Engineering (NGE) (available online with references at [http://shapiro.bsd.uchicago.edu/Table4A.CellBiochemicalActivitiesUsedinNaturalGeneticEngineering\(NGE\).html](http://shapiro.bsd.uchicago.edu/Table4A.CellBiochemicalActivitiesUsedinNaturalGeneticEngineering(NGE).html)).

NGE activities	Functions
Endo- & exonucleases	Cleave nucleic acid chains or duplexes at interior (endo) or terminal (exo) positions
DNA & RNA ligases	Splice together 3' OH and 5' phosphate ends of nucleic acid chains
DNA & RNA polymerases	Template-directed DNA and RNA polymerization; both precise and error-prone processes
Ribonucleotide and deoxyribonucleotide terminal transferases	Template-independent addition of nucleotides to the 3' OH end of a nucleic acid chain
Reverse transcriptases	Make a DNA strand complementary to an RNA template
RNA chaperones	Hold RNA molecules in position for reverse transcription or splicing
Coordinated multi-protein homologous recombination (Rec) complexes	Carry out the process of reciprocal exchange or “gene conversion” between two homologous duplexes
Coordinated multi-protein non-homologous end-joining (NHEJ) complexes	Join together the ends of two linear DNA molecules; generally involves “processing” the ends so they can be ligated together
Serine & tyrosine site-specific recombinases	Carry out reciprocal exchange between specific recombination sites by a series of single-strand DNA–protein trans-esterification reactions
Transposases and integrases	Bind to specific sequences at the end of paired duplex regions to induce transient cleavages that are subsequently ligated into a target duplex to mobilize a segment of DNA
Homing endonucleases and inteins	Site-specific duplex endonucleases; sometimes encoded by self-splicing protein domains (inteins)
Retrospllicing introns	Self-splicing type II introns capable of reverse-splicing into RNA or DNA strands

describing the constellation of activities involved in DNA (genome) biochemistry. Many of these activities are the same as ones used in DNA damage repair, but certain functions are present specifically for generating genomic novelties. The NGE toolbox is full.

It is important to note, where it has not been made explicit, that the different activities listed in Table 5 work in a sequentially coordinated fashion in multi-molecular complexes. As a consequence of this coordination, it makes no sense to think of any NGE process as occurring “accidentally,” certainly not in the way that conventional evolutionary biologists assume DNA structural changes to arise sporadically.

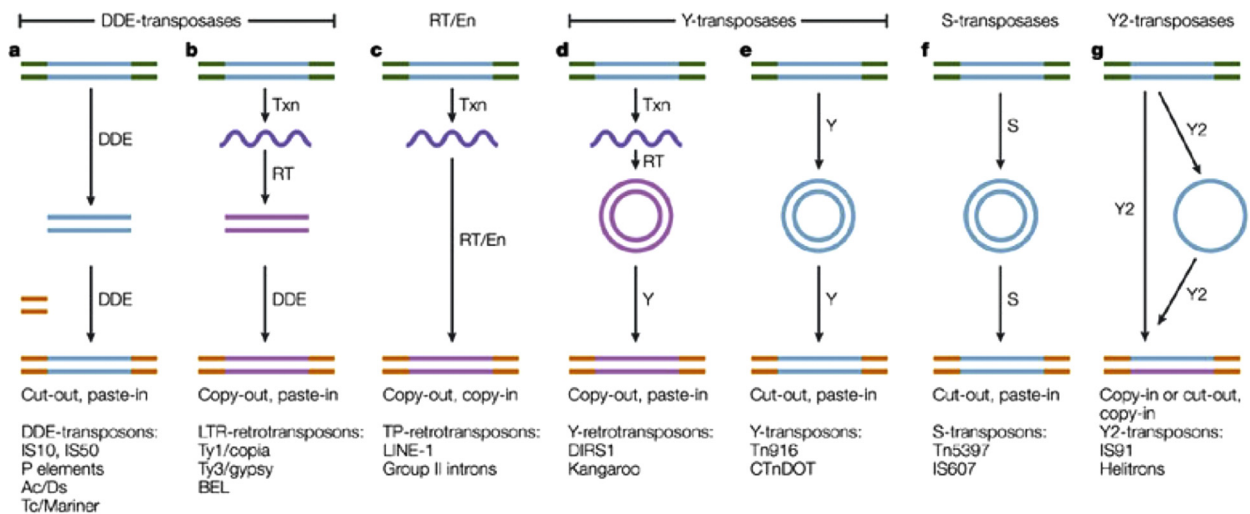


Fig. 4. Five protein families dictate different transposition pathways: DDE-transposases, reverse transcriptase/endonucleases (RT/En), tyrosine (Y)-transposases, serine (S)-transposases and rolling-circle (RC)- or Y2-transposases. Transposons (blue) can be either ‘cut-out’ or ‘copied-out’ of the flanking donor DNA (green). (a) Most DDE-transposons excise from the flanking DNA to generate an excised linear transposon, which is the substrate for integration into a target (orange). (b) Retrotransposons copy-out by reverse-transcribing (RT) a full-length copy of their RNA (purple) that is generated by transcription (Txn). Long-terminal repeat (LTR)-retrotransposons make a full-length cDNA copy (pink represents newly replicated DNA) from their RNA and integrate this into a target using a DDE-transposase. (c) TP-retrotransposons use reverse transcriptase (RT) to copy their RNA directly into a target that has been nicked by a transposon-encoded nuclease (En). (d) Y-retrotransposons are thought to generate a circular cDNA intermediate by reverse transcription. A Y-transposase integrates the element into the target. (e) and (f) Y- and S-transposons encode either a tyrosine or serine transposase, which mediates excision of the transposon to form a circular intermediate. A reversal of the catalytic steps results in transposon insertion. (g) Y2-transposons ‘paste’ one strand of the transposon into a target and use it as a template for DNA replication. Two models have been proposed for Y2-transposition. Representatives of each type of transposon are listed below each pathway [173]. Reproduced from [173] with permission from Nature Publishing Group.

## 4.2. Mobile genetic elements

The majority class of DNA in many genomes consists of dispersed repeats capable of movement from one location to another [10] (<http://shapiro.bsd.uchicago.edu/TableII.2.shtml>). These are mobile genetic elements (MGEs), originally documented in maize by McClintock in the late 1940s [171]. With the discovery of analogous elements in virtually all living cells, we can now distinguish three major classes, each of which has various subclasses [23]. In many genetic experiments, the majority of changes result from movements of MGEs. In *Drosophila melanogaster*, for example, about 85% of morphological mutations identified in the absence of mutagen treatment resulted from insertion of the *gypsy* element at new locations [172]. The movement of a defined segment of DNA carrying multiple genome formatting and coding sequences is inherently a non-random process, no matter where the insertion occurs. The best analogy is to think of the mobile element as a plug-in cassette that can be located at many sites in a circuit- or motherboard.

### 4.2.1. DNA transposons

Most mobile genetic elements found in bacteria belong to the group of DNA “transposons,” which migrate through the genome purely at the DNA level. A subset of eukaryotic mobile genetic elements also migrate exclusively as DNA [174] (Fig. 4).

DNA transposons are typically defined at each end by terminal inverted repeats (TIRs), which provide a molecular symmetry. Such symmetry is important to formation of the “transposasome” structure, involving both TIRs and multiple tightly bound copies of the “transposase” protein. The transposasome mediates DNA cleavages of the donor element and attachment to new “target sites” in the genome [23]. Intriguingly, transposases are the most common protein found in the sequence databases [175].

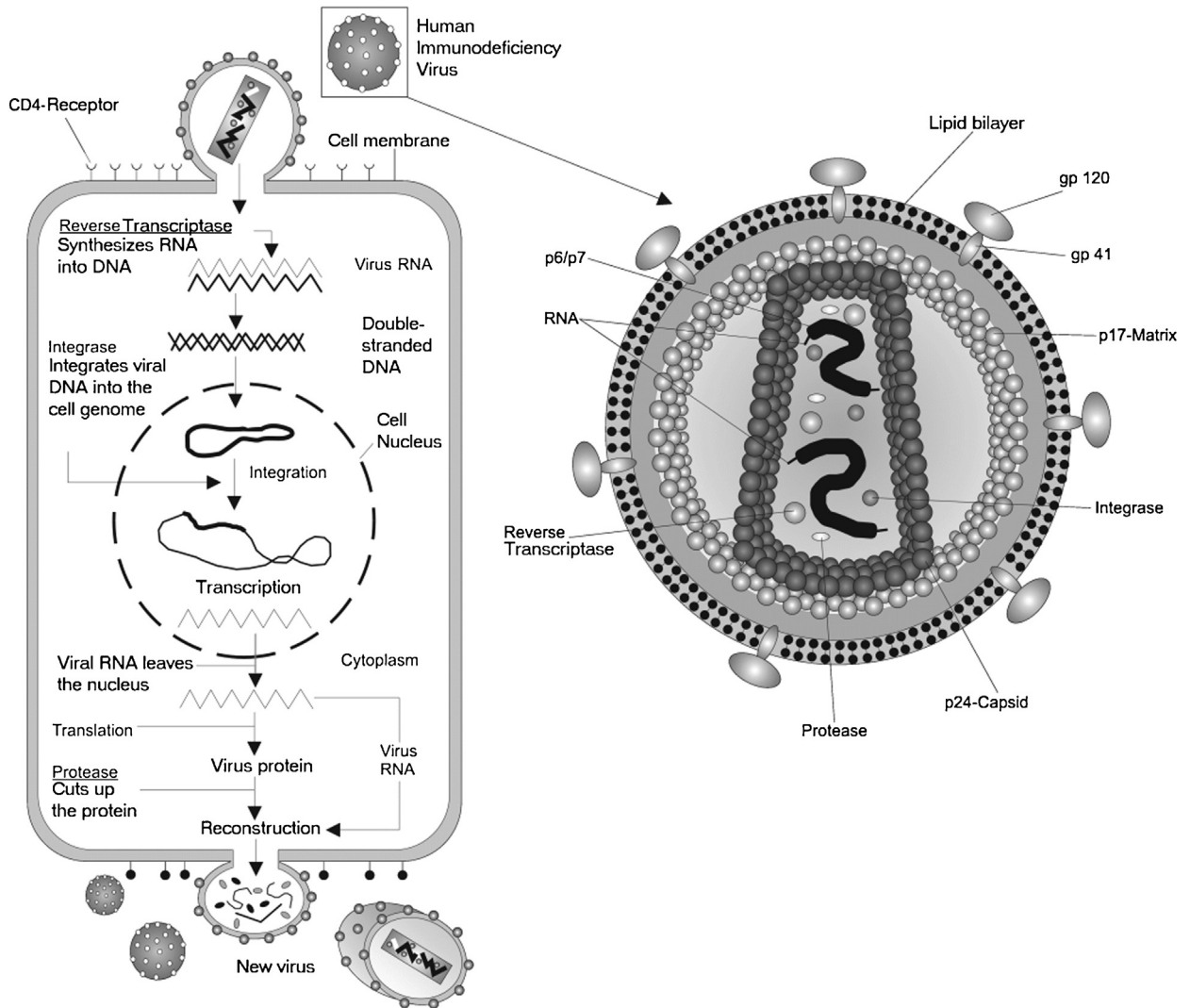


Fig. 5. HIV viral structure and reproduction cycle. Reproduced from Wikipedia with Creative Commons rights.

A remarkable variety of molecular mechanisms mobilize DNA transposons (Fig. 4). Some DNA transposons incorporate phenotypically important sequences between their termini and serve as mobilization vectors for dissemination within and between genomes. This is true for many of the antibiotic resistance coding sequences that underpin the rapid evolution of multi-drug resistance in bacteria [176]. Other examples are the exon-mobilizing “Pack-MULE” elements in the rice genome and helitrons in both plants and animals (see Table 9). These exon-carrying transposons provide us with one mechanism for the amplification and reassortment of conserved sub-protein domains that extend protein evolution beyond single amino acid changes ([http://shapiro.bsd.uchicago.edu/Exon\\_Shuffling.html](http://shapiro.bsd.uchicago.edu/Exon_Shuffling.html)).

#### 4.2.2. Retroviruses and LTR retrotransposons

Retroviruses and related elements incapable of the full viral reproduction process (retrotransposons) mobilize and amplify via RNA intermediates [177]. Retroviruses include well-known pathogens, like HIV (Fig. 5), and tumor viruses, like Rous Sarcoma Virus. Because they form infectious particles, like other transposing viruses, they can move from a donor site in one genome to a target site in another cell’s genome. Retroviral type retrotransposons can only mobilize within the genome of a single cell.

Retroviral type elements multiply with each mobilization because the donor template for the RNA intermediate remains, while a new DNA copy is produced by reverse transcription and then inserted into a target site by an “in-



tegrase” [178]. Intriguingly, the integrase operates with the same chemistry as the transposase of a transposing DNA virus [23]. These retroviral-type elements have characteristic “long terminal repeats” (LTRs) at the ends of their DNA forms and a short repeat (R) segment at the ends of the RNA intermediate. These repeated structures are integral to the complex transcription-reverse transcription process that occurs between two inserted DNA provirus stages (Fig. 6). There is also a class of terminally-repeated retrotransposons called the DIRS (direct repeat sequence) family that forms a circular cDNA copy by reverse transcription and uses a tyrosine recombinase rather than an integrase for insertion into target site DNA [179].

#### 4.2.3. *LINEs & SINEs*

There is a second class of retrotransposable elements that is not bounded by LTR repeats. These elements have been grouped under the labels “LINEs” (Long Interspersed Nucleotide Elements) and “SINEs” (Short Interspersed Nucleotide Elements) [23]. The LINE elements typically encode two proteins that mediate the reverse transcription and genome integration process, called “target-primed reverse transcription” (TPRT) (Fig. 7).

There are two major differences between the TPRT (Fig. 7) and the LTR-based retrotransposition process (Fig. 6):

- (a) TPRT starts with target site cleavage by an endonuclease activity associated with the LINE element. A 3' OH group on the cleaved target DNA serves as a primer for the first strand reverse transcription, hence the TPRT label. In LTR retrotransposition, integration into the target site comes after the full double-stranded cDNA has been produced.
- (b) In TPRT, only the 3' polyA tail of the RNA molecule is used at the start of reverse transcription. In LTR reverse transcription, priming occurs by binding a tRNA primer to a special internal “primer binding site” with the retroviral type element (Fig. 6). The lack of sequence specificity for initiating the TPRT process means that it can serve to integrate any polyA-tailed RNA molecule from the cell, including spliced mRNAs and other processed transcripts. The LINE element-encoded TPRT functions thus provide powerful tools for retaining copies of functional RNA sequences in the genome [180,181].

In many cases, LINE element transcription extends beyond the normal transcription termination/polyA addition site. Thus, sequences from the genome downstream of the LINE become incorporated into its transcript and are subsequently “retrotransduced” or inserted at a new genome location [182]. It has been demonstrated experimentally that LINE element retrotransduction of downstream sequences is a feasible mechanism of exon mobilization in protein evolution [183]. In a variation of this retrotransduction process, it has been observed that the hominid-specific SVA SINE element acquires 5' sequences as it moves to new locations, indicating that SVA transcription can begin at upstream promoters [184].

#### 4.2.4. *Mobile element formatting of target site regions*

Both McClintock and the early students of mobile elements in bacteria noted them as different because they had special and diverse effects on genome expression: novel patterns of pigment expression in maize, strong negative effects on transcription in bacteria, novel promoter activation in bacteria and yeast. These results speak to the different ways that insertion of a mobile element module can affect different stages of genome regulation (see Fig. 8 and Table 6) [24].

#### 4.3. *Activation of NGE and mobile elements by multiple stimuli*

McClintock found that introducing single broken chromosome ends into newly fertilized cells had the unexpected result of activating mobile elements that were latent and already present in her previously stable maize strains. She called such an event, which could not be repaired by pre-established routines, a “genome shock” and predicted that such shocks could lead to rapid genome restructuring (footnote 2).

There is now an extensive literature on the great diversity of challenges and stress factors that activate genome instability (<http://shapiro.bsd.uchicago.edu/TableII.7.shtml>). These factors include nutritional deprivation, intercellular signaling molecules, exposure to toxic substances (not all of which are DNA-damaging agents), and life history events such as hybridizations and infections. In eukaryotes, there is a strong correlation between these life history events and ones that disrupt epigenetic control systems (<http://shapiro.bsd.uchicago.edu/TableII.10.shtml>). The kinds

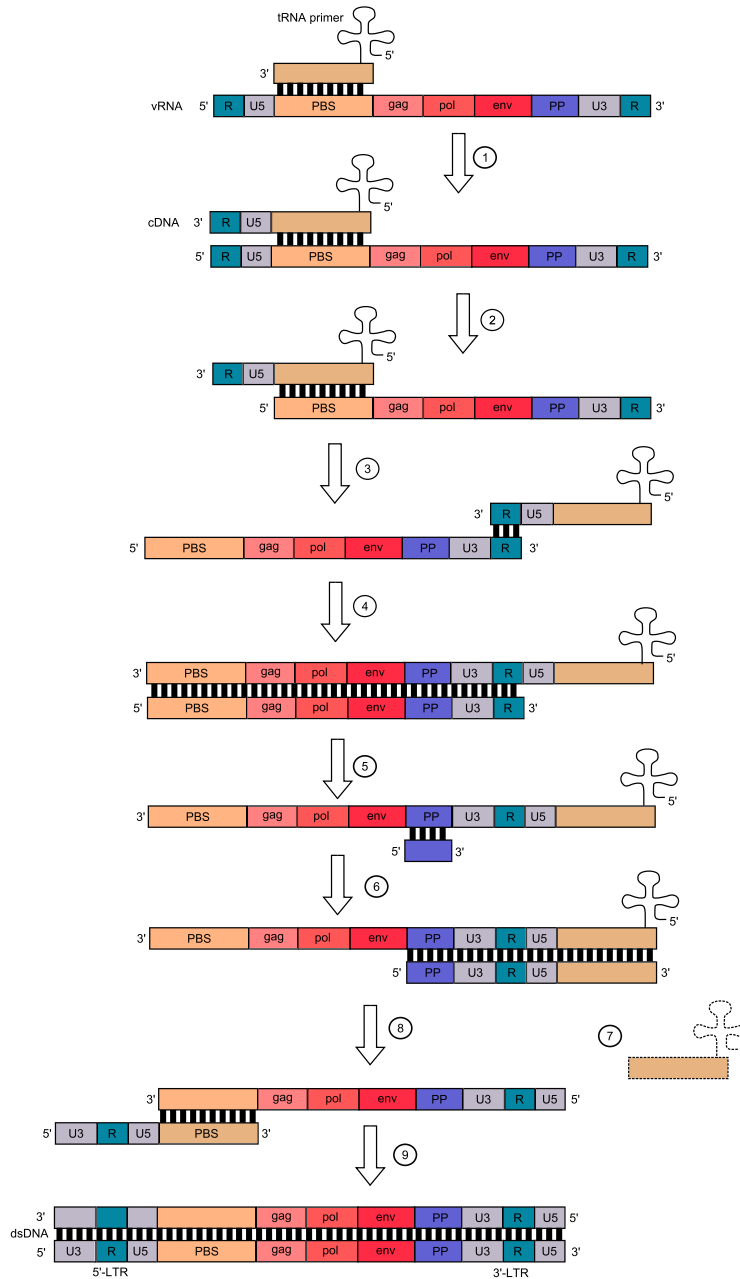


Fig. 6. Mechanism of reverse transcription in class VI virus ssRNA-RT, human immunodeficiency virus (HIV). Key: U3 – promoter region, U5 – recognition site for viral integrase; PBS – primer binding site; PP – polypurine section (polypurine tract); *gag*, *pol*, *env* – see HIV genome organization. Colors mark complementary sequences. This diagram isn't drawn to scale. Reverse transcription occurs in the cytoplasm of host cell. In this process, viral ssRNA is transcribed by the viral reverse transcriptase (RT) into double stranded DNA. Reverse transcription takes place in 5' → 3' direction. tRNA ("cloverleaf") hybridizes to PBS and provides -OH group for initiation of reverse transcription. (1) Strong stop complementary DNA (cDNA) is formed. (2) Template in RNA:DNA hybrid is degraded by RNase H domain of reverse transcriptase. (3) DNA:tRNA is transferred to the 3'-end of the template (synthesis "jumps"). (4) First strand synthesis takes place. (5) The rest of viral ssRNA is degraded by RNase H, except for PP site. (6) Synthesis of second strand of ssDNA is initiated from the 3'-end of the template. tRNA is necessary to synthesis of complementary PBS. (7) tRNA is degraded. (8) After another "jump", PBS from the second strand hybridizes with the complementary PBS on the first strand. (9) Synthesis of both strands is completed by the DNAP function of reverse transcriptase. Both dsDNA ends have U3-R-U5 sequences, so-called long terminal repeat sequences (3'-LTR and 5'-LTR, respectively). LTRs mediate integration of the retroviral DNA into another region of the host genome. Source: Alan Cann, Principles of molecular virology. Amsterdam: Elsevier Academic Press, 2005, p. 93. ISBN 0-12-088787-8. Reproduced as open access material from Wikimedia Commons.

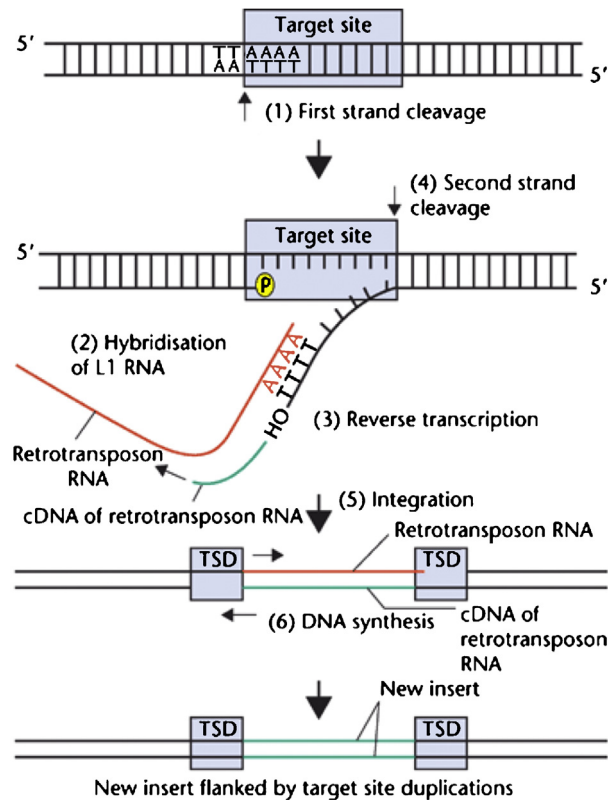


Fig. 7. Target-primed reverse transcription (TPRT). The L1 retrotransposon is thought to integrate by TPRT. (1) During L1 TPRT, the retrotransposon's endonuclease cleaves one strand of genomic DNA at the target site (grey box) cleaving at the endonuclease consensus site 5'-TT/AAAA-3': 3'-AA/TTTT-5', producing a 3' hydroxyl (OH) at the nick. (2) The retrotransposon RNA base pairs with the exposed genomic DNA strand containing Ts at the nick. (3) The retrotransposon's reverse transcriptase uses the free 3' OH to prime reverse transcription. Reverse transcription proceeds, producing a cDNA of the retrotransposon RNA. (4) A second break occurs in the other DNA strand of the target site to produce a staggered break. (5) Insertion of the cDNA into the break is completed by an unknown mechanism. (6) Removal of RNA and completion of DNA synthesis produces a complete insertion flanked by target site duplications (TSDs, grey boxes). Roy-Engel, A.M. and Belancio, V.P. 2011. Retrotransposons and Human Disease. eLS. Reproduced as open access material from Wikimedia Commons.

Table 6

More formatting of insertion targets by mobile genetic elements (available online with references at [http://shapiro.bsd.uchicago.edu/Table4B\(4\).MoreFormattingofInsertionTargetsbyMobileGeneticElements.html](http://shapiro.bsd.uchicago.edu/Table4B(4).MoreFormattingofInsertionTargetsbyMobileGeneticElements.html)).

---

Terminate transcription  
 Provide enhancer function  
 Insert splicing signals (intronization)  
 Insert coding sequences (exonization)  
 Epigenetic formatting  
 Epigenetic formatting – insulator element

---

of genome changes that occur at elevated levels in response to stress include localized point mutations, genome wide hyper-mutation, activation of mobile elements, DNA repeat losses, and chromosome rearrangements. In the bacterial case that I studied personally, the increase in frequency of CDS fusions induced by aerobic starvation could be measured as at least five orders of magnitude [185,186].

The many diverse observations on experience-activated NGE responses indicate that living organisms have the control circuitry necessary to rewrite their genomes structurally when normal reproduction is challenged. This makes sense from an evolutionary perspective because those organisms best able to change their genomes when ecological disruption occurs will be the ones most likely to leave modified progeny as survivors.

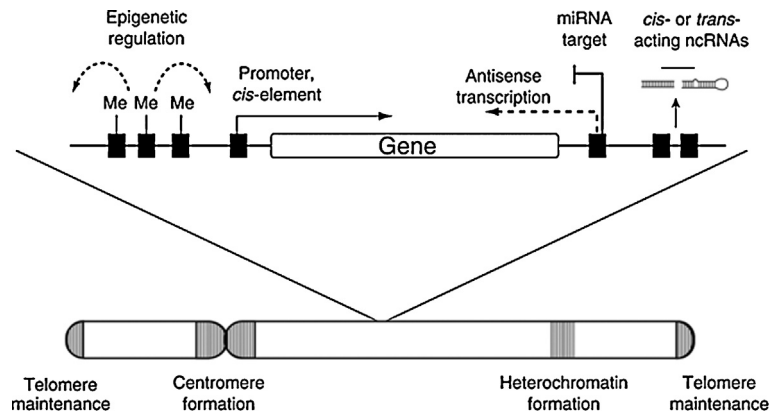


Fig. 8. Impact of transposable elements (TEs) on eukaryotic genome architecture and gene expression. TEs (shown as black vertical lines in a chromosome) can be involved in formation and maintenance of important chromosomal structures, such as centromere and telomere. TEs induce heterochromatin formation in chromosomes. TEs (shown as black boxes) can also influence local gene expression in various ways. At the transcriptional level, a TE insertion can introduce promoter sequences and *cis*-regulatory element(s) to a nearby gene. Outward-reading transcription from a TE downstream of a gene can generate an antisense RNA that potentially interfere with sense transcription. In addition, TEs can induce epigenetic gene silencing by chromatin remodeling that potentially represses the transcription of adjacent gene(s). At the post-transcriptional level, a TE that has inserted in the 3'UTR of a gene can introduce a target of miRNA that interferes translation. Finally, TEs or TE-derived sequences can produce non-coding RNAs (ncRNAs) that operate as a *cis*- or *trans*-gene regulator. Reproduced from [24] with permission from John Wiley and Sons.

#### 4.4. Targeting of NGE and mobile elements

Just as cells have the ability to silence and activate NGE functions, they also have the ability to target changes to specific sites or classes of locations in the genome [10]. There is not space to include the many examples of NGE targeting in this review, but the details are accessible in a table online (<http://shapiro.bsd.uchicago.edu/TableII.11.shtml>). The mechanisms of this targeting are all well understood in molecular biology: proteins recognizing DNA sequences, proteins recognizing DNA structures, protein–protein binding, DNA–DNA or RNA–DNA sequence interactions, and coupling to other cell functions, such as DNA replication, transcription and chromatin formatting.

Many NGE targeting modalities are of obvious adaptive utility. In terms of the spread of bacterial antibiotic resistance determinants, we can note (1) the targeting of the Tn7 transposon to replicating DNA [187], which is characteristic of plasmid molecules in the process of transferring to new cells, and (2) the specificity of integron cassette insertion for single-stranded cassettes [188], which is the molecular form in which they enter a cell during horizontal transfer by DNA uptake or conjugation. In other words, these targeting specificities are optimal for the molecular mechanisms of horizontal DNA transfer in bacteria.

In the case of yeast LTR retrotransposons, we see a different kind of adaptive utility for targeting (see Fig. 2 of [189]). The Ty1–Ty3 elements are all targeted to insert upstream of transcription start sites, which prevents them from disrupting important coding sequence data. The Ty5 element is targeted to silent chromatin regions, where it will likewise not disrupt important vital functions, except in case of stress, when its mutagenic capabilities may prove useful.

Coupling DNA restructuring to transcription is particularly important because there is no question that cells have the ability to target transcription to particular sites in the genome as part of a biologically adaptive response to external and internal circumstances. Isotype class switching in activated B cells illustrates this capability. The location of class switch recombination (CSR) is determined by lymphokine signals that control the transcription of “switch” regions where the necessary double-stranded breaks occur [190,191]. In other words, the immune system instructs the activated B cells which class of antibody to produce in response to a particular infection by targeting a specific DNA breakage-and-joining process. Thus, if we are searching for a feasible molecular mechanism to integrate widely accepted adaptive cell responses with the NGE toolbox, coordination between transcriptional regulation and genome change operations is an ideal candidate.

Table 7

Mobile genetic elements as vectors in human genetic engineering (available with references online at [http://shapiro.bsd.uchicago.edu/Table4E\(1\).MobileGeneticElementsasVectorsinHumanGeneticEngineering.html](http://shapiro.bsd.uchicago.edu/Table4E(1).MobileGeneticElementsasVectorsinHumanGeneticEngineering.html)).

Mobile element	Host organism(s)
<i>Agrobacterium</i> Ti plasmid	Plants, human cells, yeast
T DNA	
P factor transposons	<i>Drosophila melanogaster</i> and other insects
Mariner transposons	All eukaryotes
Sleeping beauty transposons	Mammals
HIV and other retroviruses	Mammals

#### 4.5. Use of NGE functions in laboratory cell manipulations

Humans take advantage of the RW nature of genomes for investigatory and practical reasons. Given the power and variety of NGE systems, it should not surprise us that they are at heart of contemporary biotechnology, cell engineering and gene therapy. Modern biotechnology utilizes enzymatic tools extracted from cells to cut, splice and mutagenize both natural and synthetic DNA: nucleases (especially restriction endonucleases), ligases, polymerases, terminal transferases, and reverse transcriptases.

##### 4.5.1. Making “knockout” and “knockin” organisms for research biotechnology and medical (gene therapy) applications

A major aspect of contemporary biotechnology is the production of deleted “knockout” organisms that have DNA removed from specific genetic loci or “transgenic” (*a.k.a.* “knockin”) organisms that have additional genetic material (“transgenes”) inserted into their germ-line genomes. A specialized but extremely important aspect of this effort is in medicine, where “gene therapy” aims at the goal of replacing defective functions or producing therapeutic proteins from introduced DNA. In all cases, following the appropriate test-tube DNA manipulations and maintenance in a microbial host, altering the target genome involves NGE functions: homologous or non-homologous recombination activities or piggy-backing on dedicated mobile element systems (Table 7).

##### 4.5.2. Synthetic biology

Another target of human inscriptions into organismal genomes is the burgeoning field known as “synthetic biology” [192]. This new area seeks to develop a lexicon of DNA elements to use as so-called “biobricks,” components that can be spliced together in different arrangements to generate regulatory and execution circuits that display predictable patterns.

Although initial synthetic biology efforts focused on transcription factors and their binding sites, a recent paper has reported the development of “biotransistors” that can execute all Boolean logical functions [193]. In terms of the RW genome perspective, it is highly significant that these biotransistor Boolean functions are executed by an NGE function: site-specific recombination organized to delete, insert, or invert specific biobricks.

#### 4.6. Epimutations over evolutionary time?

A topic of intense debate in evolutionary biology concerns the importance of heritable epigenetic changes in long-term change. The Lamarckian community has become animated about the possibility that environmentally-induced epigenetic chromatin modifications will have major phenotypic effects and will maintain themselves for sufficient generations to establish modified populations as taxonomically distinct entities [194,195]. Whether this turns out to prove the case is still undetermined. That is the reason for the query in the heading.

There are at least three sets of observations relevant to the potential evolutionary role for epigenetic chromatin modifications. Not being an expert in this field, the most I can do in this review is mention these observations and alert the reader to their potential importance in evolutionary thinking. Clearly, they fit within the RW genome perspective at shorter time scales.

Table 8

Induced transgenerational epigenetic changes (available with references online at [http://shapiro.bsd.uchicago.edu/Table4F\(1\).InducedTransgenerationalEpigeneticChanges.html](http://shapiro.bsd.uchicago.edu/Table4F(1).InducedTransgenerationalEpigeneticChanges.html)).

Stimulus or stress	Organism
Heat stress	Fission yeast ( <i>S. pombe</i> )
Heat and UV-B	<i>Arabidopsis</i>
Heavy metals	<i>Arabidopsis</i>
Heavy metals	Rice
Bacterial infection	<i>Arabidopsis</i>
Viral infection	<i>Arabidopsis</i>
Viral infection	Tobacco
Chemicals and predation	Dandelion
Predation	Milkweed
Drought exposure	<i>Polygonum persicaria</i>
Multiple stressors	Plants
RNA (induced RNAi or piRNA)	Nematode ( <i>C. elegans</i> )
Endocrine disruptors	Mice
Maternal behavior	Rats

#### 4.6.1. Induced transgenerational epigenetic effects

There is a growing literature on environmental insults which lead to transgenerational epigenetic changes. Table 8 summarizes the insults and responses. Some of these transgenerational effects can extend tens of generations.

#### 4.6.2. Spontaneous and induced epimutations

“Epimutations” is the term used to describe heritable phenotypic changes that can be mapped in crosses to specific genetic loci, which involve changes in chromatin formatting (generally detected in DNA methylation patterns) but not in DNA sequence. These arise spontaneously in plants [196], and they can also be generated predictably by passing chromosomes through genetic backgrounds deficient in DNA methylation [197]. The epimutations that form at particular loci in the absence of essential methylation activities (e.g., *FWA* = “flowering arrested” in *Arabidopsis*) persist even when reintroduced into genetic backgrounds with a full complement of methylation functions. In addition to modifying adaptive phenotypes, epigenetic changes can rapidly establish mating barriers between closely related plant populations and thus initiate independent paths of genome variation [198].

#### 4.6.3. Paramutation

One well-documented example of stable epigenetic change is “paramutation” in maize plants [199]. Paramutation refers to the phenomenon whereby certain alleles of specific genetic loci will convert, or “paramutate,” the normal wild-type form of a locus to the epimutated version in the progeny of heterozygous plants that carry both forms of the allele. No DNA sequences are altered, but epigenetic marking patterns change. These paramutations have proven stable for many generations of maize breeding. A similar form of paramutation has been reported in mice, but there is less data on how long the epimutations endure [200].

### 5. NGE and mobile elements have played an important role in genome writing during the historical record

Molecular phylogeny, developed by Woese and his colleagues in the late 1970s [201,202], has combined with whole genome sequencing to enable our understanding of cell fusions and horizontal transfer in evolutionary genome inscription (Section 2.3, see Table 3). Genome sequences further revealed the importance of whole genome duplications and other changes in chromosome number at key stages of evolutionary diversification, presumably resulting from abnormal meiosis or interspecific mating (Section 2.3). In addition, genome sequences have revealed distinct classes of repetitive and mobile DNA (<http://shapiro.bsd.uchicago.edu/TableII.2.shtml>). Thus, we know that NGE has played a major role in genome evolution, if only by contributing to establishing the major components of many genomes (<http://shapiro.bsd.uchicago.edu/TableII.1.shtml>). However, a closer look at the genome data continues to uncover important roles for mobile elements in the adaptive innovations critical to the diversification of living organisms.

Table 9

Examples of documented exon shuffling by mobile genetic elements (available with references online at <http://shapiro.bsd.uchicago.edu/Table5A.ExamplesofDocumentedExonShufflingbyMobileGeneticElements.html>).

Mobile element(s)	Organism
Mu-like element carrying exons (Pack-MULEs)	Rice
Pack-MULEs	<i>Arabidopsis</i>
Helitrons	Maize
Retrotransposon activity	Maize
Helitrons	Grass <i>Lolium perenne</i>
DNAREP1	<i>Drosophila</i>
Helitrons	<i>Drosophila</i>
Retrotransposition	<i>Drosophila</i>
Helitrons	<i>Lepidoptera</i>
LINE-1 (frequent retrotransposition from X chromosome to autosomes)	<i>Homo sapiens</i>
SVA SINE	<i>Homo sapiens</i>

### 5.1. Protein evolution by amplification and exon shuffling

Contrary to the conventional view that proteins evolve one amino acid at a time, we realized over 40 years ago that larger-scale changes play a major role. In 1970, Ohno published his book on the importance of duplications in protein diversification to assume new tasks [139]. Every taxonomic group, from bacteria to plants and animals, is characterized by its own set of protein families. Whatever the exact molecular mechanism may be, duplication of an entire protein-coding sequence can only occur by NGE, not by a succession of small mutations.

In the case of olfactory receptors, the largest protein family in mammals, there is presumptive evidence that some coding sequence copies lacking introns originated by reverse transcription [181]. Comparing the mouse and human genome complements of olfactory receptors, there is a mixed pattern of adjacent and transposed amplifications [203, 204]. Different individual olfactory receptor coding sequences have been amplified in the two genomes, indicating distinct episodes of sequence duplications to meet different adaptive needs.

In the 1970s and 1980s, we discovered (a) that protein-coding sequences are composed of discrete exons separated by introns [205,206] and (b) that protein sequences reveal an abundance of shared “domains” that define structural, functional and evolutionary modules amplified and rearranged in the course of protein evolution [207,208]. The first *Nature* paper on the draft human genome contained two figures showing “domain accretion” and “domain shuffling” in the eukaryotic evolution of transcription factors and chromatin binding proteins [21]. The sharing of domains has been an important factor in the evolution of protein interaction networks [209]. For proteins to evolve in this fashion, extended coding sequences for individual domains must be capable of both amplification (duplications) and rearrangement (cutting and splicing in new combinations).

There is a predominant but not complete correspondence between exons in the DNA sequences and the encoded protein domains [210]. Table 9 presents those organisms where genome sequence data show that coding sequence duplication and exon shuffling have occurred by NGE processes involving mobile elements.

### 5.2. Protein innovation by exonization, intronization and fusions

A key evolutionary question is, where do new coding sequences come from? Generating them one nucleotide at a time, either from existing CDSs or from non-coding sequences, would take an indeterminately long time. Moreover, there would be no reason for the intermediates to be retained by selective advantages. More rapid coding sequence innovations are essential for a plausible evolutionary mechanism. These rapid processes are known to occur through the actions of mobile elements and reverse transcription followed by genome integration.

Our recognition of how important mobile elements are as sources of coding sequences dates back to the analysis of Nekrutenko and Li (2001), when they scanned protein sequences without first removing repetitive element sequences [211]. They identified numerous protein segments that corresponded to mobile elements. Since then, the discovery of “exonized” portions of transposons and retrotransposons in various genomes has accelerated (Table 6). The same table also shows that splice sites and intron positions can be modified by mobile element insertions, thereby altering protein output, while Table 9 indicates the potential for mobile-element mediated exon shuffling.

Table 10

Reports of retrogenes in plant and animal genomes (available with references online at [http://shapiro.bsd.uchicago.edu/Table 5B. Reports of retrogenes in plant and animal genomes.html](http://shapiro.bsd.uchicago.edu/Table%205B.Reports%20of%20retrogenes%20in%20plant%20and%20animal%20genomes.html)).

Organism	Genomic results
<i>Arabidopsis</i>	15 intronized retrogenes after divergence from Papaya
Rice ( <i>Oryza sativa</i> )	27 LTR retrotransposition retrogenes 1208 primary non-LTR retrogenes, 42% chimeric; OsDR10 retrogene taxonomically restricted, encodes pathogen resistance
Poplar ( <i>Populus trichocarpa</i> )	106 retrogenes, 18 chimeric (8/18 formed after divergence from <i>Arabidopsis</i> )
Zebrafish	652 retrocopies, 440 intact retrogenes (437 transcribed), 212 pseudogenes, 10 chimerical retrogenes
Tetrapods	Ankyrin-repeat-containing <i>Sowah</i> retrogene has distinct <i>cis</i> -regulation restricted to tetrapods
Mosquitos	133 retrogenes common to <i>Aedes aegypti</i> , <i>Anopheles gambiae</i> , 57 specific to <i>A. gambiae</i>
Silkworm ( <i>Bombyx mori</i> )	68 functional inter-chromosomal retrogenes; 57% excess beyond expectation from Z sex chromosome to autosomes
<i>Drosophila</i> (X to autosomes)	Nuclear transport genes (Ntf-2 and ran) generated autosomal retroposed copies three independent times (in <i>D. melanogaster</i> , <i>D. ananassae</i> , and <i>D. grimshawi</i> lineages)
Vertebrates and insects	Retrogenes reported: 163 Human, 199 Chimp, 275 Macaca, 154 Mouse, 202 Rat, 95 Dog, 163 Cow, 232 Opossum, 99 Chicken, 140 Zebrafish, 212 Fruitfly, 108 <i>Anopheles</i>
Non-mammalian chordates	Intact retrogenes: 235 <i>Amphioxus</i> , 96 sea squirt, 151 zebrafish, 66 <i>Tetraodon</i> , 148 <i>Fugu</i> , 159 <i>Medaka</i> , 119 stickleback, 216 <i>Xenopus</i> , 217 lizard, 57 chicken, 146 <i>Platypus</i> , 565 human
Vertebrates	Retrogene origin of TM1 domain in ZIP prion proteins unique to vertebrates
Rat	26 putative TDPOZ family retrogenes in a 2.5 Mb cluster on chromosome 2
Rat	LINE-1 and ERV sequences exonized in chimeric T1, T2 TDPOZ retrogene transcripts
<i>Mus musculus</i>	Imprinted expression of four retrogenes in mouse: <i>Inpp5f_y2</i> , <i>Mcts2</i> , <i>Nap115</i> , and <i>U2af1-rs1</i>
Primates and rodents	Intronization in primate specific retrogene RNF113B; and independent “intronization” in retrogene DCAF12L2, one common to primates and rodents, one rodent-specific
<i>Homo sapiens</i>	Identified nine intron-containing retrogenes; intronizations by cryptic splice sites, silent in parental genes but active in the retrogenes
<i>Homo sapiens</i>	Up to 202 retrogenes derived from snoRNAs (small nucleolar RNAs)
<i>Homo sapiens</i>	81 chimeric retrogenes with LINE-1 signatures
<i>Homo sapiens</i>	Differential epigenetic marking of imprinted retrogenes derived from non-imprinted parental loci
<i>Homo sapiens</i>	NANOGP8 is a human-specific retro-oncogene
<i>Homo sapiens</i>	At least 18,046 “processed pseudogene” retrogene candidates identified by ENCODE
Mice and humans	Functionality of retro-“processed pseudogenes” as siRNAs and molecular decoys for regulatory RNAs controlling the parental locus ( <i>PTEN</i> and <i>KRAS</i> oncogenes)
<i>Homo sapiens</i>	Evidence that the beta-globin <i>HBBP1</i> retro-“processed pseudogene” is functional and subject to selective maintenance equal to globin coding sequences

Since mobile elements show taxonomically-specific distribution [212,213], there is great interest in using these repeats to track taxon-specific evolutionary paths [214]. In particular, exonization and domain shuffling events linked to primate-specific *Alu* SINEs and hominid-specific *SVA* SINEs have attracted great attention in the effort to discern what makes our own species unique in terms of protein-coding capacity (see Tables 6 and 10).

Suddenly arising novel coding sequences have been observed to occur by processes at both the DNA and RNA levels in diverse organisms (Table 6). Those arising by reverse transcription and genome integration (“retroposition”) presumably result from the action of LINE-encoded proteins [180,181,215] (Table 10). Many of the reverse-transcribed cDNA insertions create novel fusion sequences or “chimeric retrogenes” [216,217]. They do this both by generating novel combinations of exons and also by direct insertion of coding sequences into an existing exon.

Table 10 online also includes “pseudogenes.” These are the intronless products of reverse transcription that were thought to be without function because they carry mutations that prevent protein coding. However, we now know that pseudogenes have important regulatory effects at the post-transcriptional level [218,219].



### 5.3. Evolution of regulatory circuits by mobile elements

As mentioned previously, EvoDevo research has deepened our understanding of the role in morphogenesis played by complex multi-locus networks responding to a finite set of transcription factors [220]. Vertebrate evolution, in particular, has been viewed more as a rewiring of these networks by CRM changes rather than the evolution of new proteins [43]. By genomic analysis of morphogenetic repertoires, it has been possible to identify three discrete periods in vertebrate evolution when network amplification affected different functions [221]. “Our analysis identified three extended periods in the evolution of gene regulatory elements. Early vertebrate evolution was characterized by regulatory gains near transcription factors and developmental genes, but this trend was replaced by innovations near extracellular signaling genes, and then innovations near posttranslational protein modifiers.”

One evolutionary mystery has been how the same binding site locates at multiple dispersed genome positions fast enough to produce a useful coordinated network of the kind initially proposed by Britten and Davidson [25,222]. It would take an indefinitely long time for transcription factor binding sites, promoters, and tissue-specific enhancers to accumulate at multiple loci by independent random changes at each position. The existence of mobile genetic elements provides a mechanism for the rapid dispersal of regulatory sequences through the genome [223]. Mobile elements, such as *Alu* SINEs, contain many binding sites for transcription factors that allow them to play a role in establishing the regulation of developmental processes [224].

As predicted by many workers on transposons and retrotransposons, a very high proportion of *cis*-regulatory changes have been found to derive from mobile elements [225]. As more vertebrate genomes sequences become available, the size of the mobile element-derived regulatory site repertoire has grown significantly. According to two recent papers, the total stands at >280,000 mobile element exaptations covering ~7 MB common to mammalian genomes [226]. “Of the ~1.1 million constrained elements that arose during the 90 million years between the divergence from marsupials and the eutherian radiation, we can trace >19% to mobile element exaptations.” The proportion of regulatory elements in the human genome derived from mobile elements now stands at 11–20% of the total identified control sequences [227]. Note that these numbers represent lower bounds because the sequence traces that link sites to their mobile element origins decay over time.

It is important to bear in mind that mobile elements are taxonomically specific and so will contribute lineage-specific regulatory sequences to host genomes [228,229]. Table 11a gives a list of mobile element exaptations as regulatory sites documented in the genome sequence record. Many of these are distributed. Note that certain networks, such as the p53 and PAX6 response systems, have acquired regulatory sequences from different mobile elements in different lineages.

In addition to distributing regulatory sites to multiple genomic loci, mobile elements contribute to biological networks by neo-functionalization of transposases, the most common proteins in the sequence database [175,230,231]. Naturally, transposases have already evolved to recognize sequences in their cognate mobile elements. Table 11b provides examples of transposase exaptation (adaptation to novel function). Retroviruses and retrotransposons have similar but less extensive potential for producing novel proteins [232].

Similar but less extensive data are available showing that mobile elements and retrotransposed pseudogenes serve both as sources and targets for novel small regulatory RNAs [233–235]. Since cellular defenses against the destabilizing effects of mobile elements involves capture of DNA fragments to direct CRISPR crRNAs in prokaryotes, piRNAs in animals and analogous siRNAs in plants, there is a well-established molecular path for mobile element origins of regulatory RNAs in all domains of life [236–238].

### 5.4. Evolutionary change by altering chromosome composition and chromosome restructuring, including whole genome duplications

One major form of rapid evolutionary inscription comprises the various ways cells can change their chromosome constitution, or “karyotype” (literally, nuclear character). As discussed above in Section 2.3.3 on hybridization, one of the most common karyotype changes is duplication of the entire chromosome complement, also known as whole genome duplication (WGD) or polyploidization (Table 12).

WGD is particularly useful for the creation or exaptation of distributed genome networks by bursts of mobile element activity because the process generates a reserve copy of each network component to ensure retention of the original functionality [143]. One of the hallmarks of WGD events is the amplification of key regulatory proteins or

Table 11a

Mobile elements found to be exapted as *cis*-regulatory control sites in animals (available with references online at <http://shapiro.bsd.uchicago.edu/Table5C-1.MobileElementsFoundtoBeExaptedascis-RegulatoryControlSitesinAnimals.html>).

Organism	Locus or system	Mobile element
<i>Drosophila simulans</i>	<i>Cyp6g1</i> (encoding a cytochrome P450) – insecticide resistance	<i>Doc</i> DNA transposon
<i>Drosophila melanogaster</i>	Tissue-specific expression of the insecticide resistance gene <i>Cyp6g1</i>	<i>Cis</i> -regulatory elements in the <i>Accord</i> retrotransposon
Deuterostomes	105 highly conserved sequences from amphioxus through humans not associated with protein-coding sequences	Ancient SINE family <i>AmnSINE1</i> ( <i>Amniota SINE1</i> ), present in mammals and birds; <i>AmnSINE1</i> chimeric structure of 5S rRNA and a tRNA-derived SINE, related to five tRNA-derived SINE families in the coelacanth, dogfish shark, hagfish, and amphioxus genomes. We collectively name <i>DeuSINE</i> ( <i>Deuterostomia SINE</i> ) superfamily
Indonesian coelacanth, <i>Latimeria menadoensis</i>	Previously unknown SINE retroposon family active in <i>Sarcopterygii</i> (lobe-finned fishes and terrestrial vertebrates) in Silurian period $\geq 410$ million years ago	Mouse enhancer assay recapitulates multiple aspects of <i>Isl1</i> expression patterns. A > 200-base-pair ultraconserved region, 100% identical in mammals, and 80% identical to the coelacanth SINE, contains a 31-amino-acid-residue alternatively spliced exon of the messenger RNA processing gene <i>PCBP2</i>
Mammals (placentals, marsupials, monotremes), but not other vertebrates	Proopiomelanocortin ( <i>POMC</i> ) locus	Ancient exaptation of a <i>CORE-SINE</i> into highly conserved neuronal enhancer <i>nPE2</i>
Rodent	16 Pax6 binding sites (eye and CNS development factor)	<i>B1</i> SINE elements
<i>Homo sapiens</i>	Novel PAX6 binding sites	<i>Alu</i> SINE repeats at 3 loci
Mouse	<i>Lama3</i> transcript, stimulated by transcription factor USF (upstream stimulatory factor)	<i>B2</i> SINE elements provide RNA polymerase II promoter
Mammals	CTCF (11-zinc finger protein or CCCTC-binding transcription factor) networks; transcriptional regulation, insulator activity, V(D)J joining and chromatin formatting	Fossilized repeat elements flank conserved CTCF-binding regions, indicating retrotransposon expansions hundreds MYA; repeat-driven dispersal of CTCF binding is ancient and still highly active
Mouse	FGF8 (fibroblast growth factor 8)	<i>AmnSINE1</i> (AS071) enhancer recapitulates FGF8 expression in developing forebrain, diencephalon and hypothalamus
Mouse	SINE 390 kbp upstream of <i>Satb2</i>	<i>AmnSINE1</i> (AS021) SINE displays specific enhancer activity in developing cerebral cortex
Rodents, humans	Anti-apoptotic locus NAIP (BIRC1)	Repeated recruitment of LTR retrotransposons as promoters during mammalian evolution; 5' flanking regions of IAP family as group, in both human and mouse, enriched for LTR insertions
<i>Homo sapiens</i>	Novel NAIP isoform	<i>Alu</i> SINE element start site
Primates (baboon, human)	Primate <i>beta3GAL-T5</i> colon expression	<i>ERV1</i> LTR promoter
<i>Homo sapiens</i>	Enhancer in last intron of human <i>CD8 alpha</i> locus	<i>Alu</i> SINE-containing, T-cell-specific enhancer
<i>Homo sapiens</i>	Down syndrome critical region 4 (DSCR4) and DSCR8 loci	<i>ERV1</i> LTR acts as a bidirectional promoter
Mouse, human	REST (RE1-Silencing Transcription Factor) targets many neural genes in preimplantation embryo and CNS by cognate DNA motif, RE1 (Repressor Element 1)	1301 and 997 RE1s in human and mouse genomes, respectively, >40% are novel (many <i>LINE-1</i> , <i>LINE-2</i> , <i>Alu</i> SINE in human genome); weak RE1 elements in <i>hERV</i>
Mouse	Estrogen Receptor $\alpha$ (ER $\alpha$ ) response loci	MIR (mammalian interspersed repeat) SINEs 15–25% of TEs conserved between human and mouse; include targets ER $\alpha$ , such as GREB1, RAR $\alpha$ . TAP/SEC14L2, GLUT1/SLC2A1 (ER $\alpha$ mediated response to hypoxia), Anxa6, Fyn, CA12, CYP1B1, KRT13 and PRKACA

Table 11a (Continued.)

Organism	Locus or system	Mobile element
Zebrafish ( <i>Danio rerio</i> )	180 p53 responsive loci; human orthologs contribute neuronal morphogenesis, axonogenesis, synaptic transmission and regulation of programmed cell death	<i>EnSpmN6_DR</i> non-autonomous DNA transposon
<i>Homo sapiens</i>	p53 regulon	1509 of ~319,000 human ERV LTR regions have a near-perfect p53 DNA-binding site; LTR10 and MER61 families particularly enriched; primate-specific
<i>Homo sapiens</i> and mouse	c-Myc regulatory subnetwork	Thousands TEs bound by c-Myc; 816-4564 contain canonical c-Myc binding motifs; c-Myc binding sites over-represented among ancient families <i>LINE2</i> and <i>MIR</i> ; loci with TE-derived c-Myc binding sites co-expressed with each other and with c-Myc. Conserved fragments of <i>MIR</i> and <i>LINE2</i> transposable elements in intergenic regions within the mouse and human genomes
Eutherian mammals	Endometrial expression in placental mammals	13% of 1532 endometrial-expressed loci are within 200 kb of a Eutherian-specific MER20 transposable element; these elements have epigenetic signatures of enhancers, insulators and repressors and directly bind transcription factors essential for pregnancy to regulate expression in response to progesterone and cAMP

highly evolved structures, such as the quadrupling of Hox clusters in vertebrates and their octuplication in teleost fishes [141]. In addition, WGD events are marked by morphological innovations, like the vertebrate skeleton [239], and expansions of functional networks. Examples of the latter include MADS-box transcription factors [240] and photosynthetic functions [241] in flowering plants, globins in vertebrates [242], G protein coupled receptors in teleosts [243], and GATA-binding transcription factors among vertebrates [244].

Rapid evolutionary change also occurs by changes in chromosome structure, the process long known as “chromosomal speciation” [245–249]. Chromosome rearrangements are important in taxonomic divergence because they affect both genome expression and set up meiotic barriers to interbreeding. In some cases we can link chromosome rearrangements to the presence of mobile elements [250], where they may act either as sources of dispersed homology for recombinational exchange [251] or as active agents of chromosome breakage and rejoining [252,253].

Whatever the mechanism of chromosome rearrangements may be, we have known ever since McClintock’s pioneering studies that they are not accidental but result from concerted cell activity. The number of coordinated biochemical activities needed for joining broken ends is simply too large to constitute an accidental process: the broken ends are marked epigenetically by a special histone, moved to the same repair focus in the nucleus, and then processed for joining and ligation by one of the multi-protein NHEJ complexes.

## 6. Conclusions and outstanding issues

The preceding discussion presents a large but far from exhaustive catalogue of molecular tools that cells possess to make controlled inscriptions in their genomes over all biological time scales. The empirical evidence for biological action in hereditary genome changes has become so overwhelming that it is surprising how widespread the notion of accidental change still remains.

It is essential to keep in mind that the only constant in life is change – sometimes small and of short duration, but sometimes major and for very long periods. It would be amazing if living organisms did not have the means to use their genomes in rewritable form to cope with inevitable fluctuations in the conditions of life. At some times the genome serves as RAM memory, at other times it operates more like a flash or hard drive, but it always has to receive the inscriptions needed to adapt to changing functional needs.

Nucleoprotein complex formation, epigenetic chromatin modifications, and DNA restructuring by NGE are all now well-established phenomena. If the last 60 years of molecular genetics are any guide, it is likely that we will discover still further means cells have to store and access genomic information in the future. Hopefully, this review will initiate a process of fundamental rethinking in the reader about the nature of genome change. To facilitate that reevaluation, allow me to draw a few conclusions and point to a fundamental unanswered question.

Table 11b

Exaptation of transposases to regulatory and other protein functions (available with references only at <http://shapiro.bsd.uchicago.edu/Table5C-2.ExaptationofTransposasestoRegulatoryandOtherProteinFunctions.html>).

Organism	New protein function(s)
Budding yeast ( <i>Kluyveromyces lactis</i> )	Mating-type switch specific endonuclease
Fission yeast ( <i>Schizosaccharomyces pombe</i> )	Retrotransposon silencing
Fission yeast and mammals	Domestication of <i>pogo</i> -like transposases into centromere-binding proteins
Ciliate protozoa	DNA cleavage for macronuclear development
<i>Arabidopsis</i>	DAYSLEEPER from hAT transposase essential for normal plant growth; binds motif (Kubox1) upstream region of <i>Ku70</i> DNA repair locus
<i>Arabidopsis</i>	Transcription factors regulating light signaling
<i>Arabidopsis</i>	Maverick transposon-derived <i>Mustang</i> (MUG) family – essential for normal vigor and development
Sugarcane, rice, maize ( <i>Zea mays</i> ) and <i>Sorghum bicolor</i>	<i>Mustang</i> (MUG) family – highly conserved in grasses; ubiquitously transcribed, downregulated by phytohormones; possible involvement in hormonal homeostasis
<i>Arabidopsis</i>	<i>FHY3</i> and <i>FAR1</i> transcriptional regulatory proteins involved in phytochrome A-signaling pathway
Cereal grasses (wheat, barley, rice, oat, rye, maize, sorghum and sugarcane)	Gary protein from hAT transposase, shows conservation and positive selection across grasses but function unknown
<i>Drosophila</i> and other animals	Minos/Tc1 transposase related to paired DNA-binding domain and Pax family of transcription factors
<i>Drosophila</i>	PIF transposons protein containing a Myb/SANT domain, closely related to MADF DNA-binding domain in several <i>Drosophila</i> transcription factors
<i>Drosophila</i> and other animals	Pipsqueak (Psq) DNA-binding domain family proteins related to CENP-B/transposase subgroup
Vertebrates	HARB11, conserved protein for vertebrate development predicted nuclease, from Harbinger transposon
Vertebrates	Adaptive immune system endonuclease for VDJ joining activity in antigen receptor diversification
<i>Homo sapiens</i>	Chimeric SETMAR chromatin remodeling function from <i>Hsmar1</i> transposase
<i>Homo sapiens</i>	PogZ derived from lentivirus (retrovirus) integrase; interacts with lens epithelium-derived growth factor/p75 (LEDGF/p75)
<i>Homo sapiens</i>	Metnase/SETMAR – chromatin formatting, suppresses chromosome rearrangements, enhances DNA repair, replication, and decatenation
<i>Drosophila</i> and humans	THAP apoptosis protein DNA-binding domain related to P factor transposase
Many eukaryotes	WRKY-GCM1 Zn-finger DNA-binding domain from MULE transposases
Many eukaryotes	BED finger DNA-binding domain in chromatin insulator-binding proteins

### 6.1. Circuitry in place for controlling time and location of genome inscriptions

There is broad recognition that cells have elaborate molecular regulatory circuits and DNA formatting (Section 1) to control the timing and location of genome movements, replication, expression and repair [13]. These events occur at the shortest biological time scale. Although we know less of the details, there is a similar recognition that epigenetic chromatin formatting is subject to a comparable regulatory regime at intermediate biological time scales [254,255].

However, there is virtually no recognition that the longest-term genome inscriptions by DNA restructuring are subject to cellular control circuits. The reasons for this confusion are philosophical rather than empirical. They date back to the evolutionary debates of the late 19th and early 20th centuries. These philosophical prejudices notwithstanding, the empirical data is clear that DNA changes are equally subject to short-term and epigenetic regulation [10].

DNA changes by NGE occur in response to multiple stimuli that include (but go well beyond) DNA-damaging agents (<http://shapiro.bsd.uchicago.edu/TableII.7.shtml>). We have a growing literature on the nucleoprotein and epigenetic circuitry that regulates NGE functions. Some of these control circuits appear briefly in Sections 2, 3 and 4 of this review, and a far more extensive bibliography is available online (<http://shapiro.bsd.uchicago.edu/ExtraRefs>).

Table 12

Whole genome duplications or other amplifications observed in the genome sequence record (available with references online at <http://shapiro.bsd.uchicago.edu/Table5D.WholeGenomeDuplications.html>).

---

Yeast ( <i>Saccharomyces cerevisiae</i> )
Fungi ( <i>Rhizopus oryzae</i> , rice pathogen, and <i>Mycosphaerella graminicola</i> , wheat pathogen)
<i>Paramecium tetraurelia</i> (3 rounds of WGD)
Angiosperms
<i>Arabidopsis</i>
Rice and other cereals ( <i>Poaceae</i> )
Wheat ( <i>Triticum</i> )
Maize ( <i>Zea mays</i> )
Ragwort ( <i>Senecio</i> )
Fireweed ( <i>Chamerion angustifolium</i> )
Apple ( <i>Malus</i> )
Salsify ( <i>Tragopogon</i> , <i>Asteraceae</i> )
Vertebrates
Teleost fishes ( <i>Fugu</i> , <i>Tetraodon nigroviridis</i> )
Salmonids

---

[CellularRegulationNaturalGeneticEngineering.shtml](http://shapiro.bsd.uchicago.edu/TableII.1.1.1.shtml)). Among the important aspects of NGE regulation is the ability to target DNA change operators by well-established molecular mechanisms to regions with special distinguishing characteristics, such as sequence, structure, epigenetic formatting, replication status, or transcriptional activity (<http://shapiro.bsd.uchicago.edu/TableII.1.1.1.shtml>).

Combining the demonstrated abilities cells possess to regulate the timing and location of DNA changes, it is therefore empirically well-established that the NGE toolbox contains all the functionalities necessary for a controlled process of long-term genome restructuring.

### 6.2. NGE is a superior long-term algorithm for genome innovation

If we think about the process of generating useful biological novelties by genome change, it should be obvious that NGE processes possess several advantages over a random walk through the limitless possibilities of genome space. In terms of generating completely novel genome components, such as exons, introns and regulatory sites, the NGE examples described in Sections 2, 3 and 4 provide rapid means of producing innovations, which can be tested immediately by selection for adaptive utility. If they provide adaptive added value, they will be retained, amplified and dispersed by further cell fusion and NGE activities, as we have seen in Sections 2 and 4. If the novelties are not beneficial, they will be lost or maintained in an unselected fashion.

Once cells have novel useful functional components (*i.e.*, evolutionary inventions), the data presented in Section 4 tell us that these components are shuffled to generate combinatorial diversity in protein and cell network structures. It is obvious that combining components with established functional utility is a more reliable way to generate new activities than random modification of an existing function. Since the set of combinations is smaller than the total genome space, searches for useful new functionalities will be faster and more efficient.

One final factor which allows NGE to make rapid and large changes to genome structure is that multiple changes can occur coordinately once one or more DNA change operators has been activated. Different changes do not have to be independent events, and we have well-documented examples of widespread genome changes in a single generation [95,118,256]. This activation effect on NGE functions may be one of the reasons that hybridization is such a strong stimulus to speciation (Section 2.3.3).

### 6.3. How do cellular control circuits guide adaptive outcomes?

The great challenge for the life sciences today is to understand how all the regulatory processes and control circuits operate to make indescribably complex reproductive, repair and morphogenetic processes come out right in the face of changing circumstances. Although molecular genetics has provided us with many partial answers to this question (*e.g.*, transcription networks, checkpoints, chromatin formatting), we still have no comprehensive understanding of

biological algorithms. In this regard, we are not in a greatly different situation from early 20th Century biologists who pondered mysteries of embryonic development and regeneration [257].

However, there are two advantages we have over the scientists like Driesch. First, molecular analysis and genome sequencing have provided us with parts lists of sensory and decision-making circuits in each kind of living organism. We also have biophysical and biochemical data on how the parts interact. Second, we have the examples of cybernetic and electronic computational systems to use as examples of how such cell circuitry might function. *In silico* technology also provides us with methods to simulate and test ideas about biological control circuits.

We have no problem applying artificial intelligence ideas to modeling short-term genome inscriptions to help explain biological responses to fluctuating stimuli and damage. Similarly, we are open to applying information-processing ideas to the challenges of multicellular development, morphogenetic regulation and intermediate-term epigenetic inscriptions. If we truly accept the RW genome view, then we should be equally open to applying those same concepts to long-term biological genome restructuring in response to major ecological disruptions. The toolbox is clearly available to execute an informed process of innovative evolutionary natural genetic engineering. The time has arrived to investigate complex genome changes in real time and try to learn how NGE has worked so successfully in adapting life to the turbulent history of our planet.

## Acknowledgements

I thank Perry Miller for pointing out the mobile element-plug-in cassette analogy and Richard A Watson for letting me know that Turing's 1937 paper included Write capacity in the universal Turing machine.

## References

- [1] Lamarck J-B. *Philosophie zoologique*, original edition of 1809 with introduction by Andre Pichot. Paris: Flammarion; 1994.
- [2] Darwin C. *The variation of animals and plants under domestication*, 2 vols. New York: Organe Judd; 1868.
- [3] Weismann A. *The germ-plasm: a theory of heredity*. New York: Charles Scribner's Sons; 1893.
- [4] Brenner S. Turing centenary: life's code script. *Nature* 2012;482(7386):461. <http://www.ncbi.nlm.nih.gov/pubmed/22358811>.
- [5] Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 1953;171(4356):737–8. <http://www.ncbi.nlm.nih.gov/pubmed/13054692>.
- [6] Watson JD, Crick FH. Genetical implications of the structure of deoxyribonucleic acid. *Nature* 1953;171:964–7.
- [7] McClintock B. The significance of responses of the genome to challenge. *Science* 1984;226(4676):792–801. <http://www.ncbi.nlm.nih.gov/pubmed/15739260>.
- [8] Sonneborn TM. Beyond the gene. *Am Sci* 1949;37(1):33–59. <http://www.ncbi.nlm.nih.gov/pubmed/18109142>.
- [9] Sonneborn TM. Partner of the genes. *Sci Am* 1950;183(5):30–9.
- [10] Shapiro JA. *Evolution: a view from the 21st century*. Upper Saddle River, NJ: FT Press Science; 2011. p. 272.
- [11] Stamatoyanopoulos JA. What does our genome encode? *Genome Res* 2012;22(9):1602–11. <http://www.ncbi.nlm.nih.gov/pubmed/22955972>.
- [12] Turing AM. On computable numbers, with an application to the Entscheidungsproblem. *Proc Lond Math Soc* 1937;42:230–65.
- [13] Alberts B, et al. *Molecular biology of the cell*. New York/London: Garland Science; 2002.
- [14] Mueller-Hill B. *The lac operon: a short history of a genetic paradigm*. Berlin: de Gruyter; 1996.
- [15] Reznikoff WS. The lactose operon-controlling elements: a complex paradigm. *Mol Microbiol* 1992;6(17):2419–22. <http://www.ncbi.nlm.nih.gov/pubmed/1328815>.
- [16] Davidson EH. *The regulatory genome*. San Diego: Academic Press; 2006.
- [17] Davidson EH, Erwin DH. Gene regulatory networks and the evolution of animal body plans. *Science* 2006;311(5762):796–800. <http://www.ncbi.nlm.nih.gov/pubmed/16469913>.
- [18] Bernstein BE, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489(7414):57–74. <http://www.ncbi.nlm.nih.gov/pubmed/22955616>.
- [19] Britten R, Kohne DE. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* 1968;161:529–40. <http://www.ncbi.nlm.nih.gov/pubmed/4874239>.
- [20] Britten RJ, Kohne DE. Repeated segments of DNA. *Sci Am* 1970;222(4):24–31. <http://www.ncbi.nlm.nih.gov/pubmed/5417826>.
- [21] Lander ES, et al. Initial sequencing and analysis of the human genome. *Nature* 2001;409(6822):860–921. <http://www.ncbi.nlm.nih.gov/pubmed/11237011>.
- [22] de Koning AP, et al. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* 2011;7(12):e1002384. <http://www.ncbi.nlm.nih.gov/pubmed/22144907>.
- [23] Craig N, Craigie R, Gellert M, Lambowitz AM. *Mobile DNA II*. Washington: American Society for Microbiology Press; 2002.
- [24] Nakayashiki H. The Trickster in the genome: contribution and control of transposable elements. *Genes Cells* 2011;16(8):827–41. <http://www.ncbi.nlm.nih.gov/pubmed/21722269>.

- [25] Britten RJ, Davidson EH. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q Rev Biol* 1971;46(2):111–38. <http://www.ncbi.nlm.nih.gov/pubmed/5160087>.
- [26] Kornberg RD, Klug A. The nucleosome. *Sci Am* 1981;244(2):52–64. <http://www.ncbi.nlm.nih.gov/pubmed/7209486>.
- [27] Clark DJ. Nucleosome positioning, nucleosome spacing and the nucleosome code. *J Biomol Struct Dyn* 2010;27(6):781–93. <http://www.ncbi.nlm.nih.gov/pubmed/20232933>.
- [28] Segal E, et al. A genomic code for nucleosome positioning. *Nature* 2006;442(7104):772–8. <http://www.ncbi.nlm.nih.gov/pubmed/16862119>.
- [29] Gilbert N, Bickmore WA. The relationship between higher-order chromatin structure and transcription. *Biochem Soc Symp* 2006;73:59–66. <http://www.ncbi.nlm.nih.gov/pubmed/16626287>.
- [30] Capelson M, Corces VG. Boundary elements and nuclear organization. *Biol Cell* 2004;96(8):617–29. <http://www.ncbi.nlm.nih.gov/pubmed/15519696>.
- [31] Gerasimova TI, Byrd K, Corces VG. A chromatin insulator determines the nuclear localization of DNA. *Mol Cell* 2000;6(5):1025–35. <http://www.ncbi.nlm.nih.gov/pubmed/11106742>.
- [32] Misteli T. Concepts in nuclear architecture. *BioEssays* 2005;27(5):477–87. <http://www.ncbi.nlm.nih.gov/pubmed/15832379>.
- [33] Misteli T, Soutoglou E. The emerging role of nuclear architecture in DNA repair and genome maintenance. *Nat Rev Mol Cell Biol* 2009;10(4):243–54. <http://www.ncbi.nlm.nih.gov/pubmed/19277046>.
- [34] Mitchell JA, Fraser P. Transcription factories are nuclear subcompartments that remain in the absence of transcription. *Genes Dev* 2008;22(1):20–5. <http://www.ncbi.nlm.nih.gov/pubmed/18172162>.
- [35] Meuleman W, et al. Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. *Genome Res* 2013;23(2):270–80. <http://www.ncbi.nlm.nih.gov/pubmed/23124521>.
- [36] Shapiro JA, Sternberg Rv. Why repetitive DNA is essential to genome function. *Biol Rev* 2005;80:227–50. <http://www.ncbi.nlm.nih.gov/pubmed/15921050>.
- [37] Waki H, et al. Global mapping of cell type-specific open chromatin by FAIRE-seq reveals the regulatory role of the NFI family in adipocyte differentiation. *PLoS Genet* 2011;7(10):e1002311. <http://www.ncbi.nlm.nih.gov/pubmed/22028663>.
- [38] Friar JL, Goldman T, Perez-Mercader J. Genome sizes and the Benford distribution. *PLoS ONE* 2012;7(5):e36624. <http://www.ncbi.nlm.nih.gov/pubmed/22629319>.
- [39] Tanaka K. Regulatory mechanisms of kinetochore-microtubule interaction in mitosis. *Cell Mol Life Sci* 2013;70(4):559–79. <http://www.ncbi.nlm.nih.gov/pubmed/22752158>.
- [40] Henikoff S, Ahmad K, Malik HS. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* 2001;293(5532):1098–102. <http://www.ncbi.nlm.nih.gov/pubmed/11498581>.
- [41] Sinha B, et al. Dynamic organization of chromatin assembly and transcription factories in living cells. *Methods Cell Biol* 2010;98:57–78. <http://www.ncbi.nlm.nih.gov/pubmed/20816230>.
- [42] Dufour YS, Kiley PJ, Donohue TJ. Reconstruction of the core and extended regulons of global transcription factors. *PLoS Genet* 2010;6(7):e1001027. <http://www.ncbi.nlm.nih.gov/pubmed/20661434>.
- [43] Yokoyama KD, Thorne JL, Wray GA. Coordinated genome-wide modifications within proximal promoter *cis*-regulatory elements during vertebrate evolution. *Genome Biol Evol* 2011;3:66–74. <http://www.ncbi.nlm.nih.gov/pubmed/21118975>.
- [44] Han JS, Boeke JD. LINE-1 retrotransposons: modulators of quantity and quality of mammalian gene expression? *BioEssays* 2005;27(8):775–84. <http://www.ncbi.nlm.nih.gov/pubmed/16015595>.
- [45] Rino J, Carmo-Fonseca M. The spliceosome: a self-organized macromolecular machine in the nucleus? *Trends Cell Biol* 2009;19(8):375–84. <http://www.ncbi.nlm.nih.gov/pubmed/19616950>.
- [46] Lasda EL, Blumenthal T. Trans-splicing. *Wiley Interdiscip Rev RNA* 2011;2(3):417–34. <http://www.ncbi.nlm.nih.gov/pubmed/21957027>.
- [47] Michaeli S. Trans-splicing in trypanosomes: machinery and its impact on the parasite transcriptome. *Future Microbiol* 2011;6(4):459–74. <http://www.ncbi.nlm.nih.gov/pubmed/21526946>.
- [48] Frenkel-Morgenstern M, et al. Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts. *Genome Res* 2012;22(7):1231–42. <http://www.ncbi.nlm.nih.gov/pubmed/22588898>.
- [49] Herai RH, Yamagishi ME. Detection of human interchromosomal trans-splicing in sequence databanks. *Brief Bioinform* 2010;11(2):198–209. <http://www.ncbi.nlm.nih.gov/pubmed/19955235>.
- [50] Milanowska K, et al. RNAPATHWAYSDB – a database of RNA maturation and decay pathways. *Nucleic Acids Res* 2013;41(Database issue):268–72. <http://www.ncbi.nlm.nih.gov/pubmed/23155061>.
- [51] Brennicke A, Marchfelder A, Binder S. RNA editing. *FEMS Microbiol Rev* 1999;23(3):297–316. <http://www.ncbi.nlm.nih.gov/pubmed/10371035>.
- [52] Hood JL, Emeson RB. Editing of neurotransmitter receptor and ion channel RNAs in the nervous system. *Curr Top Microbiol Immunol* 2012;353:61–90. <http://www.ncbi.nlm.nih.gov/pubmed/21796513>.
- [53] Moller-Krull M, et al. Beyond DNA: RNA editing and steps toward Alu exonization in primates. *J Mol Biol* 2008;382(3):601–9. <http://www.ncbi.nlm.nih.gov/pubmed/18680752>.
- [54] Ptashne M, Gann A. *Genes and signals*. Cold Spring Harbor, NY: Laboratory Press; 2002.
- [55] Navarro FJ, Weston L, Nurse P. Global control of cell growth in fission yeast and its coordination with the cell cycle. *Curr Opin Cell Biol* 2012;24(6):833–7. <http://www.ncbi.nlm.nih.gov/pubmed/23182517>.
- [56] Hartwell L, Weinert TA. Checkpoints: controls that ensure the order of cell cycle events. *Science* 1989;246:629–34. <http://www.ncbi.nlm.nih.gov/pubmed/2683079>.
- [57] Weinert TA, Hartwell LH. Cell cycle arrest of *cdc* mutants and specificity of the RAD9 checkpoint. *Genetics* 1993;134(1):63–80. <http://www.ncbi.nlm.nih.gov/pubmed/8514150>.
- [58] Searle JS, et al. Proteins in the nutrient-sensing and DNA damage checkpoint pathways cooperate to restrain mitotic progression following DNA damage. *PLoS Genet* 2011;7(7):e1002176. <http://www.ncbi.nlm.nih.gov/pubmed/21779180>.

- [59] Putnam CD, Jaehnig EJ, Kolodner RD. Perspectives on the DNA damage and replication checkpoint responses in *Saccharomyces cerevisiae*. *DNA Repair* 2009;8(9):974–82. <http://www.ncbi.nlm.nih.gov/pubmed/19477695>.
- [60] Nguyen VC, et al. Replication stress checkpoint signaling controls tRNA gene transcription. *Nat Struct Mol Biol* 2010;17(8):976–81. <http://www.ncbi.nlm.nih.gov/pubmed/20639887>.
- [61] Ciapponi L, Cenci G. Telomere capping and cellular checkpoints: clues from fruit flies. *Cytogenet Genome Res* 2008;122(3–4):365–73. <http://www.ncbi.nlm.nih.gov/pubmed/19188707>.
- [62] Huen MS, Chen J. Assembly of checkpoint and repair machineries at DNA damage sites. *Trends Biochem Sci* 2010;35(2):101–8. <http://www.ncbi.nlm.nih.gov/pubmed/19875294>.
- [63] Fang SC, de los Reyes C, Umen JG. Cell size checkpoint control by the retinoblastoma tumor suppressor pathway. *PLoS Genet* 2006;2(10):e167. <http://www.ncbi.nlm.nih.gov/pubmed/17040130>.
- [64] Musacchio A. Spindle assembly checkpoint: the third decade. *Philos Trans R Soc Lond B, Biol Sci* 2011;366(1584):3595–604. <http://www.ncbi.nlm.nih.gov/pubmed/22084386>.
- [65] Caydasi AK, Ibrahim B, Pereira G. Monitoring spindle orientation: spindle position checkpoint in charge. *Cell Div* 2010;5:28. <http://www.ncbi.nlm.nih.gov/pubmed/21143992>.
- [66] Waddington CH. The epigenotype. *Endeavour* 1942;1:18–20 (reprinted 1977).
- [67] Matthews KS. DNA looping. *Microbiol Rev* 1992;56(1):123–36. <http://www.ncbi.nlm.nih.gov/pubmed/1579106>.
- [68] Jenuwein T, Allis CD. Translating the histone code. *Science* 2001;293(5532):1074–80. <http://www.ncbi.nlm.nih.gov/pubmed/11498575>.
- [69] Gardner KE, Allis CD, Strahl BD. Operating on chromatin, a colorful language where context matters. *J Mol Biol* 2011;409(1):36–46. <http://www.ncbi.nlm.nih.gov/pubmed/21272588>.
- [70] Vavouri T, Lehner B. Human genes with CpG island promoters have a distinct transcription-associated chromatin organization. *Genome Biol* 2012;13(11):R110. <http://www.ncbi.nlm.nih.gov/pubmed/23186133>.
- [71] Game JC, Chernikova SB. The role of RAD6 in recombinational repair, checkpoints and meiosis via histone modification. *DNA Repair* 2009;8(4):470–82. <http://www.ncbi.nlm.nih.gov/pubmed/19230796>.
- [72] Fnu S, et al. Methylation of histone H3 lysine 36 enhances DNA repair by nonhomologous end-joining. *Proc Natl Acad Sci USA* 2011;108(2):540–5. <http://www.ncbi.nlm.nih.gov/pubmed/21187428>.
- [73] Battu A, Ray A, Wani AA. ASF1A and ATM regulate H3K56-mediated cell-cycle checkpoint recovery in response to UV irradiation. *Nucleic Acids Res* 2011;39(18):7931–45. <http://www.ncbi.nlm.nih.gov/pubmed/21727091>.
- [74] Capurso D, Xiong H, Segal MR. A histone arginine methylation localizes to nucleosomes in satellite II and III DNA sequences in the human genome. *BMC Genomics* 2012;13:630. <http://www.ncbi.nlm.nih.gov/pubmed/23153121>.
- [75] Partridge JF, et al. cis-Acting DNA from fission yeast centromeres mediates histone H3 methylation and recruitment of silencing factors and cohesin to an ectopic site. *Curr Biol* 2002;12(19):1652–60. <http://www.ncbi.nlm.nih.gov/pubmed/12361567>.
- [76] Rogakou EP, et al. DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. *J Biol Chem* 1998;273(10):5858–68. <http://www.ncbi.nlm.nih.gov/pubmed/9488723>.
- [77] Coleman-Derr D, Zilberman D. Deposition of histone variant H2A.Z within gene bodies regulates responsive genes. *PLoS Genet* 2012;8(10):e1002988. <http://www.ncbi.nlm.nih.gov/pubmed/23071449>.
- [78] Ku M, et al. H2A.Z landscapes and dual modifications in pluripotent and multipotent stem cells underlie complex genome regulatory functions. *Genome Biol* 2012;13(10):R85. <http://www.ncbi.nlm.nih.gov/pubmed/23034477>.
- [79] Hu G, et al. H2A.Z facilitates access of active and repressive complexes to chromatin in embryonic stem cell self-renewal and differentiation. *Cell Stem Cell* 2013;12(2):180–92. <http://www.ncbi.nlm.nih.gov/pubmed/23260488>.
- [80] Zhou Z, et al. Structural basis for recognition of centromere histone variant CenH3 by the chaperone Scm3. *Nature* 2011;472(7342):234–7. <http://www.ncbi.nlm.nih.gov/pubmed/21412236>.
- [81] Sanei M, et al. Loss of centromeric histone H3 (CENH3) from centromeres precedes uniparental chromosome elimination in interspecific barley hybrids. *Proc Natl Acad Sci USA* 2011;108(33):E498–505. <http://www.ncbi.nlm.nih.gov/pubmed/21746892>.
- [82] Migicovsky Z, Kovalchuk I. Epigenetic modifications during angiosperm gametogenesis. *Front Plant Sci* 2012;3:20. <http://www.ncbi.nlm.nih.gov/pubmed/22645573>.
- [83] Sapienza C. Genome imprinting: an overview. *Dev Genet* 1995;17(3):185–7. <http://www.ncbi.nlm.nih.gov/pubmed/8565324>.
- [84] Ng RK, Gurdon JB. Epigenetic inheritance of cell differentiation status. *Cell Cycle* 2008;7(9):1173–7. <http://www.ncbi.nlm.nih.gov/pubmed/18418041>.
- [85] Watanabe A, Yamada Y, Yamanaka S. Epigenetic regulation in pluripotent stem cells: a key to breaking the epigenetic barrier. *Philos Trans R Soc Lond B, Biol Sci* 2013;368(1609):20120292. <http://www.ncbi.nlm.nih.gov/pubmed/23166402>.
- [86] Verdel A, et al. Common themes in siRNA-mediated epigenetic silencing pathways. *Int J Dev Biol* 2009;53(2–3):245–57. <http://www.ncbi.nlm.nih.gov/pubmed/19412884>.
- [87] Sugiyama T, et al. RNA-dependent RNA polymerase is an essential component of a self-enforcing loop coupling heterochromatin assembly to siRNA production. *Proc Natl Acad Sci USA* 2005;102(1):152–7. <http://www.ncbi.nlm.nih.gov/pubmed/15615848>.
- [88] Megosh HB, et al. The role of PIWI and the miRNA machinery in *Drosophila* germline determination. *Curr Biol* 2006;16(19):1884–94. <http://www.ncbi.nlm.nih.gov/pubmed/16949822>.
- [89] Juliano C, Wang J, Lin H. Uniting germline and stem cells: the function of Piwi proteins and the piRNA pathway in diverse organisms. *Annu Rev Genet* 2011;45:447–69. <http://www.ncbi.nlm.nih.gov/pubmed/21942366>.
- [90] Siomi MC, et al. PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol* 2011;12(4):246–58. <http://www.ncbi.nlm.nih.gov/pubmed/21427766>.
- [91] Brennecke J, et al. An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science* 2008;322(5906):1387–92. <http://www.ncbi.nlm.nih.gov/pubmed/19039138>.



- [92] Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. *Science* 2010;327(5962):167–70. <http://www.ncbi.nlm.nih.gov/pubmed/20056882>.
- [93] Luria SE, Delbrück M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 1943;28(6):491–551. <http://www.ncbi.nlm.nih.gov/pubmed/17247100>.
- [94] Wisniewski-Dye F, Vial L. Phase and antigenic variation mediated by genome modifications. *Antonie Van Leeuwenhoek* 2008;94(4):493–515. <http://www.ncbi.nlm.nih.gov/pubmed/18663597>.
- [95] Prescott DM. Genome gymnastics: unique modes of DNA evolution and processing in ciliates. *Nat Rev Genet* 2000;1(3):191–8. <http://www.ncbi.nlm.nih.gov/pubmed/11252748>.
- [96] Juranek SA, Lipps HJ. New insights into the macronuclear development in ciliates. *Int Rev Cytol* 2007;262:219–51. <http://www.ncbi.nlm.nih.gov/pubmed/17631190>.
- [97] Nowacki M, Shetty K, Landweber LF. RNA-mediated epigenetic programming of genome rearrangements. *Annu Rev Genomics Hum Genet* 2011;12:367–89. <http://www.ncbi.nlm.nih.gov/pubmed/21801022>.
- [98] Mollenbeck M, et al. The pathway to detangle a scrambled gene. *PLoS ONE* 2008;3(6):e2330. <http://www.ncbi.nlm.nih.gov/pubmed/18523559>.
- [99] Kazazian Jr HH. Mobile DNA transposition in somatic cells. *BMC Biol* 2011;9:62. <http://www.ncbi.nlm.nih.gov/pubmed/21958341>.
- [100] Thomas CA, Paquola AC, Muotri AR. LINE-1 retrotransposition in the nervous system. *Annu Rev Cell Dev Biol* 2012;28:555–73. <http://www.ncbi.nlm.nih.gov/pubmed/23057747>.
- [101] Perrat PN, et al. Transposition-driven genomic heterogeneity in the Drosophila brain. *Science* 2013;340(6128):91–5. <http://www.ncbi.nlm.nih.gov/pubmed/23559253>.
- [102] Margulis L, Sagan D. *Acquiring genomes: a theory of the origins of species*. Amherst, MA: Perseus Books Group; 2002.
- [103] Margulis L. The origin of plant and animal cells. *Am Sci* 1971;59(2):230–5. <http://www.ncbi.nlm.nih.gov/pubmed/5170543>.
- [104] Margulis L. *Symbiosis in cell evolution*. London: W.H. Freeman Co.; 1981.
- [105] Woese CR. On the evolution of cells. *Proc Natl Acad Sci USA* 2002;99(13):8742–7. <http://www.ncbi.nlm.nih.gov/pubmed/12077305>.
- [106] Woese CR. Endosymbionts and mitochondrial origins. *J Mol Evol* 1977;10(2):93–6. <http://www.ncbi.nlm.nih.gov/pubmed/592424>.
- [107] Embley TM, Martin W. Eukaryotic evolution, changes and challenges. *Nature* 2006;440(7084):623–30. <http://www.ncbi.nlm.nih.gov/pubmed/16572163>.
- [108] Keeling PJ. Chromalveolates and the evolution of plastids by secondary endosymbiosis. *J Eukaryot Microbiol* 2009;56(1):1–8. <http://www.ncbi.nlm.nih.gov/pubmed/19335769>.
- [109] Janouskovec J, et al. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proc Natl Acad Sci USA* 2010;107(24):10949–54. <http://www.ncbi.nlm.nih.gov/pubmed/20534454>.
- [110] Keeling PJ. The endosymbiotic origin, diversification and fate of plastids. *Philos Trans R Soc Lond B, Biol Sci* 2010;365(1541):729–48. <http://www.ncbi.nlm.nih.gov/pubmed/20124341>.
- [111] Archibald JM, et al. Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigelowiella natans*. *Proc Natl Acad Sci USA* 2003;100(13):7678–83. <http://www.ncbi.nlm.nih.gov/pubmed/12777624>.
- [112] Sakamoto W, Miyagishima SY, Jarvis P. Chloroplast biogenesis: control of plastid development, protein import, division and inheritance. *Arabidopsis Book* 2008;6:e0110. <http://www.ncbi.nlm.nih.gov/pubmed/22303235>.
- [113] Syvanen M, Kado C. *Horizontal gene transfer*. 2nd ed. London: Academic Press; 2002.
- [114] Watanabe T. Infectious drug resistance. *Sci Am* 1967;217(6):19–28. <http://www.ncbi.nlm.nih.gov/pubmed/6061177>.
- [115] Arnold ML. Transfer and origin of adaptations through natural hybridization: were Anderson and Stebbins right? *Plant Cell* 2004;16(3):562–70. <http://www.ncbi.nlm.nih.gov/pubmed/15004269>.
- [116] Anderson E, Stebbins Jr GL. Hybridization as an evolutionary stimulus. *Evolution* 1954;8:378–88.
- [117] Stebbins Jr GL. Cataclysmic evolution. *Sci Am* 1951;184(4):54–9.
- [118] Bregliano J, Kidwell M. Hybrid dysgenesis. In: Shapiro J, editor. *Mobile genetic elements*. New York: Academic Press; 1983. p. 363–410.
- [119] Ishikawa R, Kinoshita T. Epigenetic programming: the challenge to species hybridization. *Mol Plant* 2009;2(4):589–99. <http://www.ncbi.nlm.nih.gov/pubmed/19825641>.
- [120] Michalak P. Epigenetic, transposon and small RNA determinants of hybrid dysfunctions. *Heredity* 2009;102(1):45–50. <http://www.ncbi.nlm.nih.gov/pubmed/18545265>.
- [121] Arnold ML. Natural hybridization and the evolution of domesticated, pest and disease organisms. *Mol Ecol* 2004;13(5):997–1007. <http://www.ncbi.nlm.nih.gov/pubmed/15078439>.
- [122] Hudson AG, Vonlanthen P, Seehausen O. Rapid parallel adaptive radiations from a single hybridogenic ancestral population. *Proc Biol Sci* 2011;278(1702):58–66. <http://www.ncbi.nlm.nih.gov/pubmed/20685705>.
- [123] Jiao Y, et al. Ancestral polyploidy in seed plants and angiosperms. *Nature* 2011;473(7345):97–100. <http://www.ncbi.nlm.nih.gov/pubmed/21478875>.
- [124] Hegarty MJ, et al. Extreme changes to gene expression associated with homoploid hybrid speciation. *Mol Ecol* 2009;18(5):877–89. <http://www.ncbi.nlm.nih.gov/pubmed/19175502>.
- [125] Eickermann M, Ulber B, Vidal S. Resynthesized lines and cultivars of *Brassica napus* L. provide sources of resistance to the cabbage stem weevil (*Ceutorhynchus pallidactylus* (Mrsh.)). *Bull Entomol Res* 2011;101(3):287–94. <http://www.ncbi.nlm.nih.gov/pubmed/21092380>.
- [126] Hegarty MJ. Changes to gene expression associated with hybrid speciation in plants: further insights from transcriptomic studies in *Senecio*. *Philos Trans R Soc Lond B* 2008;363:3055–69.
- [127] Leitch AR, Leitch IJ. Genomic plasticity and the diversity of polyploid plants. *Science* 2008;320(5875):481–3. <http://www.ncbi.nlm.nih.gov/pubmed/18436776>.
- [128] Buggs RJ, Soltis PS, Soltis DE. Does hybridization between divergent progenitors drive whole-genome duplication? *Mol Ecol* 2009;18(16):3334–9. <http://www.ncbi.nlm.nih.gov/pubmed/19627486>.

- [129] Beaulieu J, Jean M, Belzile F. The allotetraploid *Arabidopsis thaliana*–*Arabidopsis lyrata* subsp. *petraea* as an alternative model system for the study of polyploidy in plants. *Mol Genet Genom* 2009;281(4):421–35. <http://www.ncbi.nlm.nih.gov/pubmed/19148683>.
- [130] Dehal P, Boore JL. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* 2005;3(10):e314. <http://www.ncbi.nlm.nih.gov/pubmed/16128622>.
- [131] Hufton AL, et al. Early vertebrate whole genome duplications were predated by a period of intense genome rearrangement. *Genome Res* 2008;18(10):1582–91. <http://www.ncbi.nlm.nih.gov/pubmed/18625908>.
- [132] Jaillon O, et al. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 2004;431(7011):946–57. <http://www.ncbi.nlm.nih.gov/pubmed/15496914>.
- [133] Hegarty MJ, et al. Nonadditive changes to cytosine methylation as a consequence of hybridization and genome duplication in *Senecio* (Asteraceae). *Mol Ecol* 2011;20(1):105–13. <http://www.ncbi.nlm.nih.gov/pubmed/21073590>.
- [134] Koszul R, Fischer G. A prominent role for segmental duplications in modeling eukaryotic genomes. *C R Biol* 2009;332(2–3):254–66. <http://www.ncbi.nlm.nih.gov/pubmed/19281956>.
- [135] Initiative AG. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 2000;408(6814):796–815. <http://www.ncbi.nlm.nih.gov/pubmed/11130711>.
- [136] Renny-Byfield S, et al. Independent, rapid and targeted loss of highly repetitive DNA in natural and synthetic allopolyploids of *Nicotiana tabacum*. *PLoS ONE* 2012;7(5):e36963. <http://www.ncbi.nlm.nih.gov/pubmed/22606317>.
- [137] Shapiro JA. Genome system architecture and natural genetic engineering in evolution. *Ann NY Acad Sci* 1999;870:23–35. <http://www.ncbi.nlm.nih.gov/pubmed/10415470>.
- [138] Shapiro JA. *Genome informatics: the role of DNA in cellular computations*. *Biol Theory* 2006;1:288–301.
- [139] Ohno S. *Evolution by gene duplication*. London: George Allen and Unwin; 1970.
- [140] Siegel N, et al. Comparative genomics of ParaHox clusters of teleost fishes: gene cluster breakup and the retention of gene sets following whole genome duplications. *BMC Genomics* 2007;8:312. <http://www.ncbi.nlm.nih.gov/pubmed/17822543>.
- [141] Matsunami M, Sumiyama K, Saitou N. Evolution of conserved non-coding sequences within the vertebrate Hox clusters through the two-round whole genome duplications revealed by phylogenetic footprinting analysis. *J Mol Evol* 2010;71(5–6):427–36. <http://www.ncbi.nlm.nih.gov/pubmed/20981416>.
- [142] Gillis WQ, et al. Whole genome duplications and expansion of the vertebrate GATA transcription factor gene family. *BMC Evol Biol* 2009;9(1):207. <http://www.ncbi.nlm.nih.gov/pubmed/19695090>.
- [143] Jaillon O, Aury JM, Wincker P. “Changing by doubling”, the impact of whole genome duplications in the evolution of eukaryotes. *C R Biol* 2009;332(2–3):241–53. <http://www.ncbi.nlm.nih.gov/pubmed/19281955>.
- [144] Kassahn KS, et al. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Res* 2009;19(8):1404–18. <http://www.ncbi.nlm.nih.gov/pubmed/19439512>.
- [145] Wilkins AS. *The evolution of developmental pathways*. Sinauer; 2002.
- [146] Breitbart M. Marine viruses: truth or dare. *Ann Rev Mar Sci* 2012;4:425–48. <http://www.ncbi.nlm.nih.gov/pubmed/22457982>.
- [147] Witzany G, editor. *Viruses: essential agents of life*. Springer; 2012.
- [148] Kristensen DM, et al. New dimensions of the virus world discovered through metagenomics. *Trends Microbiol* 2010;18(1):11–9. <http://www.ncbi.nlm.nih.gov/pubmed/19942437>.
- [149] Feschotte C, Gilbert C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat Rev Genet* 2012;13(4):283–96. <http://www.ncbi.nlm.nih.gov/pubmed/22421730>.
- [150] Prudhomme S, Bonnaud B, Mallet F. Endogenous retroviruses and animal reproduction. *Cytogenet Genome Res* 2005;110(1–4):353–64. <http://www.ncbi.nlm.nih.gov/pubmed/16093687>.
- [151] Claverie JM, Abergel C. Open questions about giant viruses. *Adv Virus Res* 2013;85:25–56. <http://www.ncbi.nlm.nih.gov/pubmed/23439023>.
- [152] Fischer MG, et al. Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc Natl Acad Sci USA* 2010;107(45):19508–13. <http://www.ncbi.nlm.nih.gov/pubmed/20974979>.
- [153] Colson P, Raoult D. Gene repertoire of amoeba-associated giant viruses. *Intervirology* 2010;53(5):330–43. <http://www.ncbi.nlm.nih.gov/pubmed/20551685>.
- [154] Boyer M, et al. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc Natl Acad Sci USA* 2009;106(51):21848–53. <http://www.ncbi.nlm.nih.gov/pubmed/20007369>.
- [155] Desnues C, et al. Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proc Natl Acad Sci USA* 2012. <http://www.ncbi.nlm.nih.gov/pubmed/23071316>.
- [156] Ibarra B, et al. Proofreading dynamics of a processive DNA polymerase. *EMBO J* 2009;28(18):2794–802. <http://www.ncbi.nlm.nih.gov/pubmed/19661923>.
- [157] Kunkel TA, Bebenek K. DNA replication fidelity. *Annu Rev Biochem* 2000;69:497–529. <http://www.ncbi.nlm.nih.gov/pubmed/10966467>.
- [158] Kunkel TA, Erie DA. DNA mismatch repair. *Annu Rev Biochem* 2005;74:681–710. <http://www.ncbi.nlm.nih.gov/pubmed/15952900>.
- [159] Radman M, Wagner R. The high fidelity of DNA duplication. *Sci Am* 1988;259(2):40–6. <http://www.ncbi.nlm.nih.gov/pubmed/3064293>.
- [160] Weigle JJ, Bertani G. Variations des bacteriophages conditionnees par les bacteries hotes (Variations of bacteriophage conditioned by host bacteria). *Ann Inst Pasteur* 1953;84(1):175–9. <http://www.ncbi.nlm.nih.gov/pubmed/13031249>.
- [161] Weigle JJ. Induction of mutations in a bacterial virus. *Proc Natl Acad Sci USA* 1953;39(7):628–36. <http://www.ncbi.nlm.nih.gov/pubmed/16589315>.
- [162] Calsou P, Defais M. Weigle reactivation and mutagenesis of bacteriophage lambda in *lexA*(Def) mutants of *E. coli* K12. *Mol Genet Genet* 1985;201(2):329–33. <http://www.ncbi.nlm.nih.gov/pubmed/2935711>.

- [163] Hutchinson F, Stein J. Mutagenesis of ultraviolet-irradiated lambda phage by host cell irradiation: induction of Weigle mutagenesis is not an all-or-none process. *Mol Gen Genet* 1980;177(2):207–11. <http://www.ncbi.nlm.nih.gov/pubmed/6445034>.
- [164] Witkin EM, Wermundsen IE. Targeted and untargeted mutagenesis by various inducers of SOS functions in *Escherichia coli*. *Cold Spring Harbor Symp Quant Biol* 1979;43 (Pt 2):881–6. <http://www.ncbi.nlm.nih.gov/pubmed/385231>.
- [165] Napolitano R, et al. All three SOS-inducible DNA polymerases (Pol II, Pol IV and Pol V) are involved in induced mutagenesis. *EMBO J* 2000;19(22):6259–65. <http://www.ncbi.nlm.nih.gov/pubmed/11080171>.
- [166] Shemarova IV. Signaling mechanisms of apoptosis-like programmed cell death in unicellular eukaryotes. *Comp Biochem Physiol B Biochem Mol Biol* 2010;155(4):341–53. <http://www.ncbi.nlm.nih.gov/pubmed/20117230>.
- [167] Engelberg-Kulka H, Yelin I, Kolodkin-Gal I. Activation of a built-in bacterial programmed cell death system as a novel mechanism of action of some antibiotics. *Commun Integr Biol* 2009;2(3):211–2. <http://www.ncbi.nlm.nih.gov/pubmed/19641731>.
- [168] Krieghoff-Henning E, Hofmann TG. Role of nuclear bodies in apoptosis signalling. *Biochim Biophys Acta* 2008;1783(11):2185–94. <http://www.ncbi.nlm.nih.gov/pubmed/18680765>.
- [169] Holoch PA, Griffith TS. TNF-related apoptosis-inducing ligand (TRAIL): a new path to anti-cancer therapies. *Eur J Pharmacol* 2009;625(1–3):63–72. <http://www.ncbi.nlm.nih.gov/pubmed/19836385%22>.
- [170] Song J. EMT or apoptosis: a decision for TGF-beta. *Cell Res* 2007;17(4):289–90. <http://www.ncbi.nlm.nih.gov/pubmed/17426696>.
- [171] McClintock B. *Discovery and characterization of transposable elements: the collected papers of Barbara McClintock*. New York: Garland; 1987.
- [172] Green MM. Mobile DNA elements and spontaneous gene mutation. In: *Eukaryotic transposable elements as mutagenic agents*. 1988. p. 41.
- [173] Curcio MJ, Derbyshire KM. The outs and ins of transposition: from Mu to Kangaroo. *Nat Rev Mol Cell Biol* 2003;4(11):865–77. <http://www.ncbi.nlm.nih.gov/pubmed/14682279>.
- [174] Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* 2007;41:331–68. <http://www.ncbi.nlm.nih.gov/pubmed/18076328>.
- [175] Aziz RK, Breitbart M, Edwards RA. Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res* 2010;38(13):4207–17. <http://www.ncbi.nlm.nih.gov/pubmed/20215432>.
- [176] Bukhari AI, Shapiro JA, Adhya SL, editors. *DNA insertion elements, plasmids and episomes*. Cold Spring Harbor, New York: Cold Spring Harbor Press; 1977.
- [177] Coffin JM, Hughes SH, Varmus HE. *Retroviruses*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1997.
- [178] Varmus H. Reverse transcription. *Sci Am* 1987;257(3):56–64. <http://www.ncbi.nlm.nih.gov/pubmed/2443971>.
- [179] Piednoel M, et al. Eukaryote DIRS1-like retrotransposons: an overview. *BMC Genomics* 2011;12:621. <http://www.ncbi.nlm.nih.gov/pubmed/22185659>.
- [180] Brosius J. The contribution of RNAs and retroposition to evolutionary novelties. *Genetica* 2003;118:99–116.
- [181] Brosius J. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* 1999;238:115–34. <http://www.ncbi.nlm.nih.gov/pubmed/10570990>.
- [182] Pickeral OK, et al. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res* 2000;10(4):411–5. <http://www.ncbi.nlm.nih.gov/pubmed/10779482>.
- [183] Moran JV, DeBerardinis RJ, Kazazian Jr HH. Exon shuffling by L1 retrotransposition. *Science* 1999;283(5407):1530–4. <http://www.ncbi.nlm.nih.gov/pubmed/10066175>.
- [184] Damert A, et al. 5'-transducing SVA retrotransposon groups spread efficiently throughout the human genome. *Genome Res* 2009;19(11):1992–2008. <http://www.ncbi.nlm.nih.gov/pubmed/19652014>.
- [185] Shapiro JA. Observations on the formation of clones containing araB-lacZ cistron fusions. *Mol Gen Genet* 1984;194(1–2):79–90. <http://www.ncbi.nlm.nih.gov/pubmed/6233472>.
- [186] Shapiro JA. Genome organization, natural genetic engineering and adaptive mutation. *Trends Genet* 1997;13(3):98–104. <http://www.ncbi.nlm.nih.gov/pubmed/9066268>.
- [187] Peters JE, Craig NL. Tn7 recognizes transposition target structures associated with DNA replication using the DNA-binding protein TnsE. *Genes Dev* 2001;15(6):737–47. <http://www.ncbi.nlm.nih.gov/pubmed/11274058>.
- [188] Hall RM. Integrons and gene cassettes: hotspots of diversity in bacterial genomes. *Ann NY Acad Sci* 2012;1267(1):71–8. <http://www.ncbi.nlm.nih.gov/pubmed/22954219>.
- [189] Levin HL, Moran JV. Dynamic interactions between transposable elements and their hosts. *Nat Rev Genet* 2011;12(9):615–27. <http://www.ncbi.nlm.nih.gov/pubmed/21850042>.
- [190] Dunnick WA, et al. The role of germline promoters and I exons in cytokine-induced gene-specific class switch recombination. *J Immunol* 2011;186(1):350–8. <http://www.ncbi.nlm.nih.gov/pubmed/21131417>.
- [191] Kinoshita K, et al. Target specificity of immunoglobulin class switch recombination is not determined by nucleotide sequences of S regions. *Immunity* 1998;9(6):849–58. <http://www.ncbi.nlm.nih.gov/pubmed/9881975>.
- [192] Schwille P, Diez S. Synthetic biology of minimal systems. *Crit Rev Biochem Mol Biol* 2009;44(4):223–42. <http://www.ncbi.nlm.nih.gov/pubmed/19635039>.
- [193] Bonnet J, et al. Amplifying genetic logic gates. *Science* 2013. <http://www.ncbi.nlm.nih.gov/pubmed/23539178>.
- [194] Sano H. Inheritance of acquired traits in plants: reinstatement of Lamarck. *Plant Signal Behav* 2010;5(4):346–8. <http://www.ncbi.nlm.nih.gov/pubmed/20118668>.
- [195] Jablonka E, Lamb M. *Evolution in four dimensions*. Boston: MIT Press; 2005.
- [196] Eichten SR, et al. Heritable epigenetic variation among maize inbreds. *PLoS Genet* 2011;7(11):e1002372. <http://www.ncbi.nlm.nih.gov/pubmed/22125494>.
- [197] Kinoshita Y, et al. Control of FWA gene silencing in *Arabidopsis thaliana* by SINE-related direct repeats. *Plant J* 2007;49(1):38–45. <http://www.ncbi.nlm.nih.gov/pubmed/17144899>.

- [198] Durand S, et al. Rapid establishment of genetic incompatibility through natural epigenetic variation. *Curr Biol* 2012;22(4):326–31. <http://www.ncbi.nlm.nih.gov/pubmed/22285031>.
- [199] Chandler VL. Paramutation's properties and puzzles. *Science* 2010;330(6004):628–9. <http://www.ncbi.nlm.nih.gov/pubmed/21030647>.
- [200] Cuzin F, Rassoulzadegan M. Non-Mendelian epigenetic heredity: gametic RNAs as epigenetic regulators and transgenerational signals. *Essays Biochem* 2010;48(1):101–6. <http://www.ncbi.nlm.nih.gov/pubmed/20822489>.
- [201] Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci USA* 1977;74(11):5088–90. <http://www.ncbi.nlm.nih.gov/pubmed/270744>.
- [202] Woese CR, Fox GE. The concept of cellular evolution. *J Mol Evol* 1977;10(1):1–6. <http://www.ncbi.nlm.nih.gov/pubmed/903983>.
- [203] Young JM, et al. Different evolutionary processes shaped the mouse and human olfactory receptor gene families. *Hum Mol Genet* 2002;11(5):535–46. <http://www.ncbi.nlm.nih.gov/pubmed/11875048>.
- [204] Niimura Y, Nei M. Comparative evolutionary analysis of olfactory receptor gene clusters between humans and mice. *Gene* 2005;346:13–21. <http://www.ncbi.nlm.nih.gov/pubmed/15716120>.
- [205] Chambon P. Split genes. *Sci Am* 1981;244(5):60–71. <http://www.ncbi.nlm.nih.gov/pubmed/6262906>.
- [206] Gilbert W. Why genes in pieces? *Nature* 1978;271:501. <http://www.ncbi.nlm.nih.gov/pubmed/622185>.
- [207] Doolittle RF, Bork P. Evolutionarily mobile modules in proteins. *Sci Am* 1993;269(4):50–6. <http://www.ncbi.nlm.nih.gov/pubmed/8235550>.
- [208] Doolittle RF. The multiplicity of domains in proteins. *Annu Rev Biochem* 1995;64:287–314. <http://www.ncbi.nlm.nih.gov/pubmed/7574483>.
- [209] Cancherini DV, Franca GS, de Souza SJ. The role of exon shuffling in shaping protein–protein interaction networks. *BMC Genomics* 2010;11 Suppl 5:S11. <http://www.ncbi.nlm.nih.gov/pubmed/21210967>.
- [210] Kaessmann H, et al. Signatures of domain shuffling in the human genome. *Genome Res* 2002;12(11):1642–50. <http://www.ncbi.nlm.nih.gov/pubmed/12421750>.
- [211] Nekrutenko A, Li WH. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet* 2001;17(11):619–21. <http://www.ncbi.nlm.nih.gov/pubmed/11672845>.
- [212] Sternberg Rv, Shapiro JA. How repeated retroelements format genome function. *Cytogenet Genome Res* 2005;110:108–16. <http://www.ncbi.nlm.nih.gov/pubmed/16093662>.
- [213] Jurka J, et al. Distinct groups of repetitive families preserved in mammals correspond to different periods of regulatory innovations in vertebrates. *Biol Direct* 2012;7(1):36. <http://www.ncbi.nlm.nih.gov/pubmed/23098210>.
- [214] Jurka J, Bao W, Kojima KK. Families of transposable elements, population structure and the origin of species. *Biol Direct* 2011;6:44. <http://www.ncbi.nlm.nih.gov/pubmed/21929767>.
- [215] Kazazian Jr HH. Genetics. L1 retrotransposons shape the mammalian genome. *Science* 2000;289(5482):1152–3. <http://www.ncbi.nlm.nih.gov/pubmed/10970230>.
- [216] Buzdin A, et al. The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination. *Nucleic Acids Res* 2003;31(15):4385–90. <http://www.ncbi.nlm.nih.gov/pubmed/12888497>.
- [217] Buzdin AA. Retroelements and formation of chimeric retrogenes. *Cell Mol Life Sci* 2004;61(16):2046–59. <http://www.ncbi.nlm.nih.gov/pubmed/15316654>.
- [218] Muro EM, Mah N, Andrade-Navarro MA. Functional evidence of post-transcriptional regulation by pseudogenes. *Biochimie* 2011;93(11):1916–21. <http://www.ncbi.nlm.nih.gov/pubmed/21816204>.
- [219] Tutar Y. Pseudogenes. *Comp Funct Genomics* 2012;2012:424526. <http://www.ncbi.nlm.nih.gov/pubmed/22611337>.
- [220] Carroll SB. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 2008;134(1):25–36. <http://www.ncbi.nlm.nih.gov/pubmed/18614008>.
- [221] Lowe CB, et al. Three periods of regulatory innovation during vertebrate evolution. *Science* 2011;333(6045):1019–24. <http://www.ncbi.nlm.nih.gov/pubmed/21852499>.
- [222] Britten RJ. DNA sequence insertion and evolutionary variation in gene regulation. *Proc Natl Acad Sci USA* 1996;93(18):9374–7. <http://www.ncbi.nlm.nih.gov/pubmed/8790336>.
- [223] Feschotte C. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* 2008;9(5):397–405. <http://www.ncbi.nlm.nih.gov/pubmed/18368054>.
- [224] Polak P, Domany E. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics* 2006;7:133. <http://www.ncbi.nlm.nih.gov/pubmed/16740159>.
- [225] Chenais B, et al. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene* 2012;509(1):7–15. <http://www.ncbi.nlm.nih.gov/pubmed/22921893>.
- [226] Lindblad-Toh K, et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 2011;478(7370):476–82. <http://www.ncbi.nlm.nih.gov/pubmed/21993624>.
- [227] Lowe CB, Haussler D. 29 Mammalian genomes reveal novel exaptations of mobile elements for likely regulatory functions in the human genome. *PLoS ONE* 2012;7(8):e43128. <http://www.ncbi.nlm.nih.gov/pubmed/22952639>.
- [228] Marino-Ramirez L, et al. Transposable elements donate lineage-specific regulatory sequences to host genomes. *Cytogenet Genome Res* 2005;110(1–4):333–41. <http://www.ncbi.nlm.nih.gov/pubmed/16093685>.
- [229] Mikkelsen TS, et al. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* 2007;447(7141):167–77. <http://www.ncbi.nlm.nih.gov/pubmed/17495919>.
- [230] Piriyaopongsa J, et al. Evaluating the protein coding potential of exonized transposable element sequences. *Biol Direct* 2007;2:31. <http://www.ncbi.nlm.nih.gov/pubmed/18036258>.
- [231] Alzohairy AM, et al. Transposable elements domesticated and neofunctionalized by eukaryotic genomes. *Plasmid* 2013;69(1):1–15. <http://www.ncbi.nlm.nih.gov/pubmed/22960324>.
- [232] Volf JN. Cellular genes derived from Gypsy/Ty3 retrotransposons in mammalian genomes. *Ann NY Acad Sci* 2009;1178:233–43. <http://www.ncbi.nlm.nih.gov/pubmed/19845640>.

- [233] Piriyaopongsa J, Jordan IK. Dual coding of siRNAs and miRNAs by plant transposable elements. *RNA* 2008;14(5):814–21. <http://www.ncbi.nlm.nih.gov/pubmed/18367716>.
- [234] Smalheiser NR, Torvik VI. Alu elements within human mRNAs are probable microRNA targets. *Trends Genet* 2006;22(10):532–6. <http://www.ncbi.nlm.nih.gov/pubmed/16914224>.
- [235] Tam OH, et al. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* 2008;453(7194):534–8. <http://www.ncbi.nlm.nih.gov/pubmed/18404147>.
- [236] Brennecke J, et al. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* 2007;128(6):1089–103. <http://www.ncbi.nlm.nih.gov/pubmed/17346786>.
- [237] Karginov FV, Hannon GJ. The CRISPR system: small RNA-guided defense in bacteria and archaea. *Mol Cell* 2010;37(1):7–19. <http://www.ncbi.nlm.nih.gov/pubmed/20129051>.
- [238] Girard A, Hannon GJ. Conserved themes in small-RNA-mediated transposon control. *Trends Cell Biol* 2008;18(3):136–48. <http://www.ncbi.nlm.nih.gov/pubmed/18282709>.
- [239] Zhang G, Cohn MJ. Genome duplication and the origin of the vertebrate skeleton. *Curr Opin Genet Dev* 2008;18(4):387–93. <http://www.ncbi.nlm.nih.gov/pubmed/18721879>.
- [240] Veron AS, Kaufmann K, Bornberg-Bauer E. Evidence of interaction network evolution by whole-genome duplications: a case study in MADS-box proteins. *Mol Biol Evol* 2007;24(3):670–8. <http://www.ncbi.nlm.nih.gov/pubmed/17175526>.
- [241] Coate JE, Doyle JJ. Divergent evolutionary fates of major photosynthetic gene networks following gene and whole genome duplications. *Plant Signal Behav* 2011;6(4):594–7. <http://www.ncbi.nlm.nih.gov/pubmed/21494088>.
- [242] Hoffmann FG, Opazo JC, Storz JF. Whole-genome duplications spurred the functional diversification of the globin gene superfamily in vertebrates. *Mol Biol Evol* 2011. <http://www.ncbi.nlm.nih.gov/pubmed/21965344>.
- [243] Semyonov J, et al. GPCR genes are preferentially retained after whole genome duplication. *PLoS ONE* 2008;3(4):e1903. <http://www.ncbi.nlm.nih.gov/pubmed/18382678>.
- [244] Gillis WQ, et al. Whole genome duplications and expansion of the vertebrate GATA transcription factor gene family. *BMC Evol Biol* 2009;9:207. <http://www.ncbi.nlm.nih.gov/pubmed/19695090>.
- [245] White MJ. Chromosomes of the vertebrates. *Evolution* 1949;3(4):379–81. <http://www.ncbi.nlm.nih.gov/pubmed/15396755>.
- [246] White MJD. *Animal cytology and evolution*. Cambridge University Press; 1945. 4th ed. 1973.
- [247] King M. *Species evolution: the role of chromosome change*. Cambridge University Press; 1995.
- [248] Voss SR, et al. Origin of amphibian and avian chromosomes by fission, fusion, and retention of ancestral chromosomes. *Genome Res* 2011;21(8):1306–12. <http://www.ncbi.nlm.nih.gov/pubmed/21482624>.
- [249] Ruiz-Herrera A, Farre M, Robinson TJ. Molecular cytogenetic and genomic insights into chromosomal evolution. *Heredity* 2012;108(1):28–36. <http://www.ncbi.nlm.nih.gov/pubmed/22108627>.
- [250] Lonnig WE, Saedler H. Chromosome rearrangements and transposable elements. *Annu Rev Genet* 2002;36:389–410. <http://www.ncbi.nlm.nih.gov/pubmed/12429698>.
- [251] Robberecht C, et al. Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. *Genome Res* 2013;23(3):411–8. <http://www.ncbi.nlm.nih.gov/pubmed/23212949>.
- [252] Lim JK, Simmons MJ. Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. *BioEssays* 1994;16(4):269–75. <http://www.ncbi.nlm.nih.gov/pubmed/8031304>.
- [253] Engels WR, Preston CR. Formation of chromosome rearrangements by P factors in *Drosophila*. *Genetics* 1984;107(4):657–78. <http://www.ncbi.nlm.nih.gov/pubmed/6086453>.
- [254] Bonisch C, et al. Chromatin proteomics and epigenetic regulatory circuits. *Expert Rev Proteomics* 2008;5(1):105–19. <http://www.ncbi.nlm.nih.gov/pubmed/18282127>.
- [255] Stein GS, et al. Architectural genetic and epigenetic control of regulatory networks: compartmentalizing machinery for transcription and chromatin remodeling in nuclear microenvironments. *Crit Rev Eukaryot Gene Expr* 2010;20(2):149–55. <http://www.ncbi.nlm.nih.gov/pubmed/21133844>.
- [256] Engels WR. P elements in *Drosophila*. *Curr Top Microbiol Immunol* 1996;204:103–23. <http://www.ncbi.nlm.nih.gov/pubmed/8556863>.
- [257] Driesch H. *The science and philosophy of the organism: Gifford lectures delivered at Aberdeen university, 1907–1908 by Hans Driesch, vol. 1*. Aberdeen: Aberdeen University; 1908.