GBE

Genome Biology and Evolution

SMBE

Widespread Recurrent Evolution of Genomic Features

Ignacio Maeso[1], Scott William Roy[2,3], and Manuel Irimia[2,4,*]

[1]Department of Zoology, University of Oxford, United Kingdom

[2]Department of Biology, Stanford University

[3]Department of Biology, San Francisco State University

[4]Banting and Best Department of Medical Research, Donnelly Centre, University of Toronto, Canada

All authors contributed equally to this work.

*Corresponding author: E-mail: mirimia@gmail.com.

Accepted: 6 March 2012

Abstract

The recent explosion of genome sequences from all major phylogenetic groups has unveiled an unexpected wealth of cases of recurrent evolution of strikingly similar genomic features in different lineages. Here, we review the diverse known types of recurrent evolution in eukaryotic genomes, with a special focus on metazoans, ranging from reductive genome evolution to origins of splice-leader trans-splicing, from tandem exon duplications to gene family expansions. We first propose a general classification scheme for evolutionary recurrence at the genomic level, based on the type of driving force—mutation or selection—and the environmental and genomic circumstances underlying these forces. We then discuss various cases of recurrent genomic evolution under this scheme. Finally, we provide a broader context for repeated genomic evolution, including the unique relationship of genomic recurrence with the genotype–phenotype map, and the ways in which the study of recurrent genomic evolution can be used to understand fundamental evolutionary processes.

Key words: genome, convergence, parallel evolution, genotype–phenotype map.

Evolutionary Biology in the Era of Ubiquitous Genomes

The explosion of genomic sequences over the past few years has revolutionized our understanding of evolution. Ten years after publication of the human genome sequence (Lander et al. 2001; Venter et al. 2001), hundreds of genomes are now available, spanning nearly all major phylogenetic groups, and providing an increasingly focused picture of evolutionary processes. These resources have allowed identification of troves of both broadly shared genomic features (allowing the reconstruction of presumed ancestral traits, e.g., the gene complements of the eukaryotic and metazoan ancestors; Putnam et al. 2007; Fritz-Laylin et al. 2010) and lineage-specific genomic changes (in some cases allowing associations with phenotypic novelties, e.g., Wang et al. 2005; Zhang et al. 2010; McLean et al. 2011). In addition, many instances of a third more puzzling phylogenetic pattern have been observed: traits whose distribution is "scattered" across the evolutionary tree (fig. 1), indicating repeated independent evolution of similar genomic features in different lineages.

Recurrent Evolution: Phenotypic, Molecular, and Genomic

Recurrent evolution has been extensively studied at a variety of levels and has often led to confusion due to a lack of explicit definitions (Doolittle 1994; Arendt and Reznick 2008). It is therefore useful to begin our discussion by comparing recurrent genomic evolution as defined and reviewed here with previous definitions and work.

Recurrent Phenotypic Evolution

Recurrent evolution has most commonly been studied at the level of organismal phenotype (fig. 2), comprising an extremely rich field with hundreds of articles spanning three centuries exploring a wide diversity of recurrent phenotypes and lineages (Scotland 2011). A central concern of phenotypic work has been understanding the physical or genetic causes for recurrence. This pursuit often focuses on distinguishing between convergent evolution and parallel evolution (a distinction which itself has been extensively debated; Arendt and Reznick 2008; Scotland 2011). Generally, the distinctions follow etymology: parallel comes from the

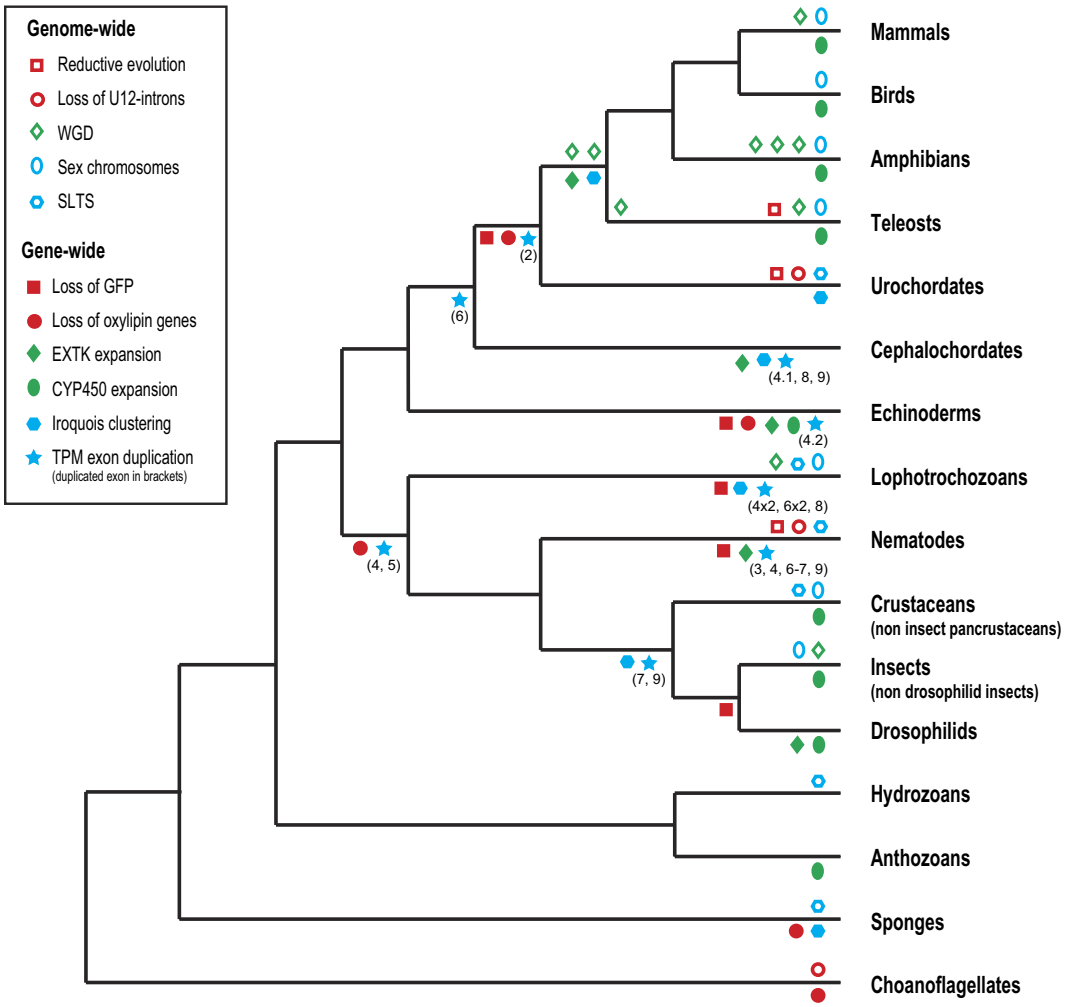

Fig. 1.—Phylogenetic distribution of some genomic features across metazoans. Genome-wide/gene-wide traits are mapped to a phylogenetic tree of metazoans (plus choanoflagellates) depicted by empty/solid forms above/below the tree branches, as indicated in the legend. Red shapes denote recurrent loss of ancestral features, whereas green features involve overall gain of genomic sequence; blue represents more complex characters. Each symbol indicates that a particular feature has evolved independently at least once within the corresponding taxonomic group. For example, "reductive evolution" in the teleost branch indicates that at least one lineage within the group (pufferfish) is known to show this feature. In the case of WGD, several symbols along the same branch represent the existence of lineages with successive rounds of WGD (i.e., octoploidy, dodecaploidy, etc.). Numbers in parentheses indicate which tropomyosin (TPM) exon(s) have duplicated in tandem in each event. The cases represented here are selected examples from the literature and are not intended as an exhaustive list; in addition, many yet unknown cases are expected to be discovered with the increasing availability of whole-genome sequences.

Greek for "beside" + "each other" (Παρά + άλληλος) and thus involves lineages with initially similar starting points arriving at similar endpoints by taking similar paths; on the other hand, convergence comes from the Latin for "with or together" (*com*-) and "to incline, tend toward" (*vergere*) and thus generally involves lineages with different starting points taking different paths to arrive at similar endpoints. For instance, one proposed distinction between parallelism and convergence focuses on the starting points for the two lineages: whether similar (closely related species, parallel) or different (distantly related species, convergent). Another proposed distinction focuses on paths (the specific genetic mutations underlying the changes) taken by the two lineages—whether the same (parallel) or different (convergent) (Arendt and Reznick 2008). Importantly, the two proposed distinctions are related since, because of their higher genetic and developmental similarities, closely related species are more likely to evolve similar traits by identical genetic changes than are species with more disparate biology (although this is not always the case; Arendt and Reznick 2008).

Recurrent Molecular Evolution

An equally diverse range of phenomena is subsumed under the heading of "recurrent molecular evolution." A useful starting point here is Doolittle's (1994) four-category

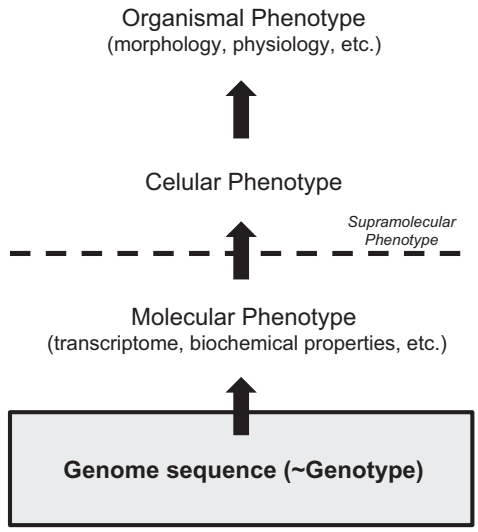

Fig. 2.—Levels of recurrent evolution. Different levels of biological organization in which recurrent evolution may be studied. Although the phenotype should be considered a continuum across the different scales of biological complexity, for practical reasons, we may divide it into three levels: 1) organismal: individual features such as anatomy, physiology, behavior, etc.; 2) cellular: characteristics of single cells, including cell movements, secretory capacities, morphology, organellar composition, etc. (equivalent to the organismal level in unicellular species); and 3) molecular: all observed traits below the cellular level, including transcriptome, proteome, biochemical properties, chromatin structure, etc. Genomic level (gray box) corresponds only to the nucleotide sequence (i.e., elements that can be recognized at the sequence level) and may be comparable to the classic concept of genotype.

schema. He identified 1) functional convergence, in which the same molecular function arises multiple times (e.g., unrelated enzymes catalyzing the same reaction; Galperin and Koonin 2012); 2) mechanistic convergence, involving evolution of similar mechanisms for accomplishing similar functions in unrelated molecules (e.g., similar sidechain geometries in unrelated serine proteases; Kraut 1977); 3) structural convergence, in which unrelated sequences fold into similar structures (e.g., repeated evolution of alpha-helices and beta-sheets or similar RNA secondary structures); and 4) sequence convergence, in which similar specific molecular (either DNA or protein) sequences evolve multiple times independently.

Recurrent Genomic Evolution

We are now in a position to define recurrent genomic evolution, the topic of this review, and to see how it differs from nearly all of these other levels of recurrence. Whereas organismal (i.e., anatomical, physiological, etc.) and most categories of molecular recurrence are observed at any of the phenotypic levels (fig. 2), genomic recurrence is directly observed as similar changes in the genotype—that is, at the level of DNA sequence. Notably, then, even most of Doolittle's molecular categories (functional, mechanistic, and structural) do not qualify as genomic recurrence because they relate to phenotype. Although these categories are defined at the molecular level (and thus intuitively "closer" to the genomic or genotypic level), they are in fact phenotypic. This becomes clear when the general definition of phenotype—the observable characteristics of an organism—is recalled. We may recognize different "levels" across the phenotypic continuum—molecular, cellular, and organismal (fig. 2)—but this does not change the fact that they are all clearly aspects of phenotype and not genotype: they reflect directly observable characteristics of the organism or cell.

Another fundamental distinction between "classical" and genomic recurrence involves the focus of the study: in classical studies of recurrence in molecules, cells, or organisms, repeated evolution is initially observed at the phenotypic level and only thereafter interrogated at the genotype/genomic level. By contrast, genotypic convergence involves direct observation of similar or same changes in the genome in different lineages, notwithstanding these changes' effects on the various levels of phenotype (whether similar, different, or even potentially nonexistent). Genotypic (~genomic) recurrence is thus most closely related to Doolittle's fourth category, sequence convergence.

The Importance of Being Recurrent and the "Rules" of Evolution

The study of recurrent evolution is of special importance for understanding the forces shaping genomes. Because of the inherent stochasticity of evolutionary processes, inferring evolutionary forces from the occurrence of a given (set of) change(s) in a single lineage is difficult. Recurrent evolution of the same genomic characteristics suggests predictability of evolution, elucidating the rules of genome evolution by revealing commonalities of evolutionary forces experienced across disparate lineages (Conway Morris 2009). We believe that the wealth of recurrent genomic features indicate unappreciated similarity of fundamental forces across lineages. Although the large number of genomic characters and finite nature of sequence space implies that genomic recurrence may sometimes occur simply by chance (see below), many cases have now been unearthed that suggest specific forces driving genome evolution down similar paths in different lineages. Identifying and understanding these forces or causes are perhaps the major challenge of the study of recurrent genome evolution.

Chance, Heterogeneity of Causes, and Genomic Recurrence

Inherent to the treatment of recurrence as a valuable and biologically meaningful tool to understand evolution is the notion that cases of repeated genomic evolution are informative if they occur in excess of the level of coincidence

expected simply from the action of stochastic processes in finite sequence space. In some cases discussed here, this null hypothesis can be rejected. Other cases await direct testing, generally because of the lack of enough data to assess the statistical significance of the pattern and/or to properly define the null hypothesis (i.e., specific mutation rates across lineages, etc.). Although we have chosen to discuss mostly cases that we believe are likely to reflect unexpected levels of recurrence (with some exceptions such as whole-genome duplications [WGDs], see below), it remains possible that some of these examples do not significantly differ from the chance expectation. Similarly, it is worth pointing out that different instances of a particular trait may be due to different pressures acting in different lineages (this is particularly possible for cases in which fundamentally different mechanisms for a given genomic change are imaginable). Although recurrent patterns caused by different pressures should be considered true recurrence, their subsequent evolutionary interpretation will be much more obscure. These considerations place similar caveats on most or all cases discussed below, and thus, they will not be discussed extensively for each instance, but just in a few particularly enlightening examples. Ultimately, random chance and our proposed explanations represent testable alternative hypotheses that could and should be directly tested.

The Causes of Recurrent Evolution of Genomic Features

What forces may explain genomic recurrence? In contrast to recurrent anatomical or physiological characters, which are usually (and reasonably) assumed to reflect adaption, often due to shared peculiarities of the organisms' environmental niches, the potential causes of observed recurrent genomic features are more diverse and may be very different for different recurrent traits—indeed, in some cases, the adaptative value of repeated genomic outcomes is dubious. In understanding the forces driving recurrent genomic evolution, we believe that the following two axes are particularly important.

Forces Driving the Pressure: Mutation, Positive Selection, or Relaxed Selection

A species undergoes a genomic change when 1) a spontaneous mutation occurs and 2) the resultant mutated allele spreads through the population, a process highly dependent on selective strength and efficiency (incorporating demography, effective population size, etc.). Thus, insofar as recurrent changes reflect similar pressures or constraints across lineages, these similarities may involve forces that are "mutational" or "selective" (or even both). The notion that selection could impart a directionality to evolutionary change is familiar to any evolutionary biologist; however, that mutation could be directional may be less familiar (the interested reader should consult Yampolsky and Stoltzfus 2001). Mutation can be no less a directional force if a certain class of mutation (G-to-A, small genomic deletions, intron loss, etc.) is more frequent than its reverse (A-to-G, small insertions, intron gain, etc.). Thus, all that is needed for mutation-driven recurrent evolution is that multiple lineages are experiencing similar mutational biases in parallel.

For selective pressure, a second question is whether the recurrence is due to similar "positive" selective pressure in multiple lineages or to similar "relaxation" of selective pressure in multiple lineages. Notably, differences in selective pressure include not only classical fitness variation but also in effective population size (N_e) that leads to differences in the effectiveness of selection versus drift. Indeed, according to one influential model, a general prediction of this is that several general aspects of the genome architecture should evolve recurrently in lineages exposed long enough to similar N_e (and mutation rates) (Lynch and Conery 2003; Lynch 2006, 2007).

Nature of the Pressure: General, Recurrent Environmental, or Recurrent Genetic

Another important consideration involves the distribution of the pressure driving convergence and the source of that pressure. Similar evolutionary pressures and constraints in two lineages can either be 1) "general" (or ancestral), that is, applying to most or all lineages within a group or 2) "recurrent," that is, pressures that themselves arose independently in only a subset of lineages. For recurrent pressure, a second question is whether the pressure arose due to a previous change in the genome of the species ("genetic" or intrinsic) or in its environment ("environmental" or extrinsic).

Using this framework, we next review some of the major known cases (or classes) of recurrent genomic evolution (summarized in table 1), beginning with the illustrative case of reductive genome evolution (RGE). Notably, for many of the phenomena discussed here, the causes remain unclear and often debated. Our goal is to frame the questions and to engender debate, not to arbitrate between competing hypotheses. In addition, we have chosen to focus on eukaryotic nuclear genomes, and thus, we will not discuss an equal number of interesting cases of recurrent evolution in prokaryotes and eukaryotic organelles.

An Example: On the Causes of Reductive Genome Evolution

These distinctions are illustrated by different hypotheses about the evolutionary causes of RGE. RGE is perhaps the best-known instance of recurrent genome evolution. RGE has been observed in nearly all eukaryotic superkingdoms (Venkatesh et al. 2000; Lane et al. 2007; Morrison et al. 2007; Opperman et al. 2008; Slamovits and Keeling 2009; Ankarklev et al. 2010; Corradi et al. 2010) and can

Table 1
Possible Causes of Recurrent Genomic Evolution

	Driving force			Nature of the pressure			Probability of occurrence by chance
		Selectional			Recurrent		
	Mutational	Positive	Relaxed	General/Ancestral	Environmental	Genetic	
Genomic organization							
Reductive evolution	X	X	X		X	X	Null
Genome expansion	X	X	X			X	Low
WGDs	X			X			High
Sex chromosomes	X					X	Low
Nucleotide composition	X	X				X	Low
Genome-wide gene structures							
Massive intron loss	X	X			X	X	Low
Strong intron boundaries		X				X	Null
SLTS	X					X	Low
Complete loss of ancestral U12 introns	X			X			Low
Gene/gene family level							
Gene family expansions		X		X	X		High
Cluster formation and assembly of syntenic blocks		X					Low
Disruption of gene clusters and other syntenic blocks			X			X	High
Gene losses	X		X		X	X	High
Specific intragenic features							
Tandem exon duplications		X		X			Low
Gene structures	X			X			High
Loss of gene segments			X			X	Low

include pronounced gene loss, elimination of repetitive elements, evolution of overlapping genes, reduction of average intron sizes, and/or intron numbers and other genomic changes leading to more compact genomes. In addition, significant genome contractions have occurred even in typically large genomes: For instance, multiple mammalian orders have experienced parallel patterns of genome contraction (including loss of nuclear mitochondrial sequences [NumtS], pseudogenes, and long terminal repeat retrotransposons) following the Cretaceous–Tertiary (KT) boundary (Rho et al. 2009).

Several hypotheses have been proposed for genome reduction. First, RGE is often argued to reflect positive selection for loss of inessential genomic elements acting specifically on parasitic/fast-replicating lineages. This hypothesis is an example of a recurrent (acting only or especially on some lineages) environmental (due to considerations of an organism's niche) "positive-selective" pressure. Another alternative is that RGE reflects loss of genomic sequences that are no longer efficiently maintained by selection ("relaxed-selective" pressure). Several possible reasons for relaxed-selective pressure are possible. Changes in lifestyle could render some processes obsolete (e.g., parasites that obtain products from their hosts may lose biosynthetic pathways), an example of "recurrent-environmental" causes. Reduced efficiency of selection due to reduced effective population size in parasites could also lead to weakly selected elements (also recurrent-environmental) (Lynch 2007). In some cases, loss of one gene may render related/interacting genes nonfunctional, leading to their loss. This case of relaxed-selective pressure is due to changes within the organism's genome (gene loss) and thus is a case of recurrent-genetic. Finally, it is also possible that some aspects of RGE simply reflect a strong tendency toward deletion at the genome level (mutational pressure). Such a deletion process could arise due to changes in the DNA replication/repair machinery (genetic) or due to changes in the environment (e.g., increased ultra violet exposure leading to a greater rate of double-strand breaks in DNA; environmental). Notably, it is also conceivable that the pressures governing recurrent RGE are general: Gene loss is known even in species without striking genome reduction, and many lineages appear to experience an excess of DNA deletions over insertion (Petrov 2002a, 2002b). From this perspective, lineages undergoing RGE could potentially be exhibiting general pressures that have simply proceeded to a more advanced stage.

Multiple Levels of Recurrent Genomic Evolution

We next proceed to a discussion of different examples of observed genomic recurrence. We have organized these examples by the "scale" of their changes: recurrent genomic evolution can be recognized at multiple scales, ranging from

whole-genome patterns such as RGE (globally affecting numerous individual features at the same time) to specific changes within individual genes (such as the recurrent deletion of a regulatory DNA motif). Although these different levels are interconnected and, in many cases, are probably interdependent, for clarity, we will divide the examples discussed here into four broad categories. We will first review cases of genome-wide patterns of recurrent evolution, subdivided into changes in genomic organization (such as RGE) and global changes of gene structures. Then, we will focus on cases affecting single genes or gene families. Last, we will zoom in to discuss examples of recurrent evolution of features within the individual genes themselves.

Cases of Recurrent Evolution of Genomic Organization

Expansive Genome Evolution

Another repeatedly observed evolutionary trajectory is pronounced expansion of genome size and content. At least in animal, plant, and fungi, some species have dramatically increased total DNA content (Gregory et al. 2007). In some cases, gene numbers have increased several-fold relative to related lineages (often through WGDs [see below]), accompanied by evolution of large gene families, apparently increased intergenic and intron lengths, and, in nearly all cases, massive proliferation of repetitive elements (e.g., Lander et al. 2001; Bennetzen 2002; Kidwell 2002; Piegu et al. 2006; Ungerer et al. 2006; Gregory et al. 2007). Similar histories may have also been experienced by other lineages; however, systematic undersampling of large genomes outside of these three groups has hampered our knowledge of other such taxa. Here, again, the causes for convergent genome expansion remain unclear, although, given that massive genome expansions require hundreds of mutations accumulating in the same direction, they are unlikely to evolve simply by chance. Some hypotheses closely associate genome expansion with multicellularity. One possibility is that multicellularity promotes evolution of regulatory complexity and gene family expansion (Vogel and Chothia 2006; Taft et al. 2007; Lang et al. 2010). Another influential hypothesis suggests that genome expansion in multicellular organisms largely reflects reduced selection against mildly deleterious insertions (such as gene duplicates, transposable element insertions, and introns) in species with reduced N_e, such as plants or animals (Lynch and Conery 2003; Lynch 2007). However, recent work questioning the correlation between N_e and genomic complexity urge caution (Whitney and Garland 2010, but see Lynch 2011). Finally, it is possible that genetic changes, such as high expression of active retrotranscriptases, can lead to increased proliferation of repeated elements, a recurrent-genetic mutational cause.

Whole-Genome Duplications

A polyploid is a cell, organism, or species that contains more than two homologous sets of chromosomes. The mutation that produce them is referred to as WGD or polyploidization, and it has been repeatedly described in many eukaryotic groups, including animals (Bisbee et al. 1977; Amores et al. 1998; Gallardo et al. 1999; Evans et al. 2004; Edger and Pires 2009), plants (Fawcett et al. 2009), ciliates (Aury et al. 2006), oomycetes (Martens and Van de Peer 2010), and fungi (Wolfe and Shields 1997; Ma et al. 2009). Although extensive gene losses in paleopolyploids could result in a diploid-like gene complement, WGDs are generally not reversible and therefore are a case of mutational ratchet, a "general mutational" cause (see below). In some lineages, this phenomenon is especially pervasive, with a high prevalence of multiple extra rounds of polyploidizations after a first WGD event (especially common in plants, but also several animal lineages) (Evans et al. 2004). However, it is not clear whether recurrent WGDs, although very frequent, occur and accumulate more often than expected for a random process. From a selectional perspective, although WGDs can have immediate phenotypic effects (Kennedy et al. 2006; Thompson and Merg 2008), these may not explain the fixation in most cases. However, Fawcett et al. (2009) have suggested that plant lineages that underwent WGDs had a better chance to survive after the KT mass extinction. In addition, WGDs have been postulated to have served as a frequent source of increased evolutionary potential for subsequent evolution (Blomme et al. 2006; Zhang and Cohn 2008), even though hypotheses linking WGDs with big taxonomic radiations and evolutionary novelties have been controversial (Donoghue and Purnell 2005; Hurley et al. 2007). In total then, although WGD may result in dramatic recurrent patterns at a genome-wide level, it may not be caused by common evolutionary forces acting on a particular set of lineages but may simply respond to a high mutational frequency (i.e., a higher rate of mutations leading to polyploidization).

Sex Chromosomes

In many distantly related eukaryotes, sex is determined at the genetic level by chromosomal complement. This is thought to involve a cascade of events driven largely by sexual antagonistic selection, including 1) a gene at a previously autosomal locus develops a dominant ability to determine sex; 2) recombination is suppressed at this locus; 3) additional sex-related genes accumulate nearby on the chromosome, further driving recombination suppression; 4) stepwise degradation of the chromosome containing the dominant sex determinant (Y/W); and 5) increased traffic of genes between the sex chromosomes and autosomes. Evolution of similar sex chromosome systems has occurred repeatedly in vertebrates, invertebrates, fungi, and plants

(Fraser et al. 2004; Fraser and Heitman 2005; Bergero et al. 2007; Bellott et al. 2010; Charlesworth and Mank 2010; Davis and Thomas 2010; Kaiser and Bachtrog 2010; Ellegren 2011). Sex chromosomes are thus an example of a "selectional" cascade of events triggered by recurrent genetic changes. Finally, another interesting case of recurrent evolution of a genome-based sex determination system is the X-autosome balance in at least *Drosophila* and *Caenorhabditis* (reviewed in Haag 2005) and the plant genus *Rumex* (Navajas-Pérez et al. 2005).

Changes in Global and Local Nucleotide Composition

Global nucleotide composition (or GC content) ranges widely across eukaryotic and prokaryotic genomes. In particular, many divergent lineages have recurrently evolved highly AT-rich genomes throughout eukaryotic evolution (Gardner et al. 2002; Eichinger et al. 2005; Eisen et al. 2006; Ghedin et al. 2007), whereas the evolution of highly GC-rich genomes is rarer among eukaryotes (Merchant et al. 2007). These differences are likely due to a combination of selectional and mutational pressures (including mutational bias and biased recombination-associated DNA repair) (Yampolsky and Stoltzfus 2001; Birdsell 2002). Interestingly, because genome-wide GC-content is a major determinant of global codon bias (Hershberg and Petrov 2009), independent evolution of similar GC-contents in two different species will usually result in recurrent evolution of similar preferential codon usages.

The same pressures—especially local differences in recombination (Duret 2006; Duret and Arndt 2008)—are likely to cause local differences in GC-content also within genomes (e.g., isochores). Notably, these regions are continuously evolving; for example, several mammalian lineages are undergoing a recurrent process of GC-rich isochore erosion, with a significant trend of G/C to A/T substitutions, whereas others are independently increasing their overall GC-content (Duret et al. 2002; Belle et al. 2004; Romiguier et al. 2010). Interestingly, in addition to repeated patterns of nucleotide composition at a genomic scale, these trends sometimes result in cases of striking recurrence of GC-content at specific genes (e.g., the gene *RAG1* in two marsupial species; Gruber et al. 2007).

Cases of Genome-Wide Recurrent Evolution of Gene Structures

Widespread Genome-Wide Intron Loss

Whereas most studied eukaryotic species have plentiful spliceosomal introns (at least one per gene on average), several distantly related lineages contain far fewer (<0.1 per gene, Matsuzaki et al. 2004; Vanacova et al. 2005; Morrison et al. 2007), apparently due to independent episodes of massive intron loss (Irimia and Roy 2008). Why should this be? Perhaps, the leading hypothesis is that massive intron reduction reflects strong positive selection for intron loss in lineages that are optimized for fast replication (Doolittle 1978). This is a recurrent-environmental positive-selection model, since it invokes increased positive selection due to peculiarities of species' environments, related to RGE. On the other hand, massive reduction in intron number could reflect "runaway" mutation, for instance due to elevated rates of creation of intronless DNA copies of genes by widespread retroposition associated with retroelement invasion (Roy and Penny 2007). This is a recurrent-genetic mutational model, since it invokes increased mutation due to peculiarities of species' genomes (retroelement invasion). Finally, evidence for more gradual intron number reduction in many lineages suggests a general mutational pressure toward intron loss, potentially due to a near absence of intron gain in many lineages (Roy and Irimia 2009a). This hypothesis provides an example of a "ratchet-like" effect (Covello and Gray 1993; Doolittle 1998), in which transition in one direction (from intron presence to absence) occurs much more readily than the reverse (intron gain), leading to a strong directionality to evolution. Ratchets can be due to mutation, selection, or a complicated combination of the two and are a common phenomenon across recurrent evolution of genomic features (see below for further discussion on the role of ratchet processes on the evolution of genome complexity and the constructive neutral evolution [CNE]; Stoltzfus 1999; Gray et al. 2010; Doolittle et al. 2011; Speijer 2011).

Transformation of Intron Structures after Massive Intron Loss

In each case in which a eukaryotic lineage has experienced nearly complete intron loss, the few remaining introns exhibit modified splicing signals, with strengthened consensus sequences for core splicing motifs (5′ splice site and branch point), and even highly constrained distance between the branch point and the 3′ intron boundary (Irimia et al. 2007, 2009; Irimia and Roy 2008; Schwartz et al. 2008). Such a tight association between two genomic transformations—intron loss and intron sequence change—suggests that genetic changes associated with one lead to selective pressures driving the other: a case of recurrent genetic positive-selective pressures. However, although several mechanistic hypotheses have been proposed (Irimia and Roy 2008; Irimia et al. 2009), a clear explanation is still lacking.

Spliced Leader Trans-Splicing

Spliced leader trans-splicing (SLTS) is a variation on the spliceosomal splicing mechanism that attaches short *trans*-encoded RNA "leader" sequences to the 5′ end of transcripts of a generally well-defined subset of genes. SLTS systems exhibit a highly punctate phylogenetic distribution

across protists and animals (Lukes et al. 2009; Roy and Irimia 2009b; Douris et al. 2010; fig. 1). Phylogenetic evidence suggests frequent evolution of SLTS from a non-SLTS ancestor; by contrast, no case of loss of SLTS in any lineage is known (Roy and Irimia 2009b), although, with current data and methods for detecting SLTS, cases of secondary loss of SLTS are hard to prove. This suggests a model in which 1) new SLTS systems arise at some rate over evolutionary time, likely by creation from spliced leader-like sequences from traditional spliceosomal RNAs by largely neutral mutations (Lukes et al. 2009) and 2) degradation of defunct 5′ untranslated regions (UTRs) following the evolution of SLTS leads to a very low probability of loss of SLTS. Thus, SLTS may be another case of mutational ratchet in which transition from one state to another is common over evolutionary time, but the reverse is rare, therefore leading to recurrent evolution of the same feature. Interestingly, the cascade of events leading to the evolution of SLTS may result in increased molecular complexity, by enabling new molecular paths of gene expression.

One instance of the increased molecular complexity associated with SLTS is the evolution of polycistronic transcripts, which is tightly associated with SLTS in diverse eukaryotic lineages (and is very rare in eukaryotes without SLTS). This difference likely reflects the fact that in eukaryotes, translation of downstream *open reading frames (*ORFs) is generally inefficient. As such, in eukaryotes that lack SLTS, polycistronic transcripts will be rare; however, SLTS upstream of ORFs can create monocistronic mature messenger RNAs from polycistronic transcripts, resolving this difficulty. Dynamics of operon creation and loss may also reflect a ratchet: Mutations affecting transcription termination of upstream genes and leading to long transcripts may allow effective expression of trans-spliced downstream genes from polycistronic messages; on the other hand, internal promoters in operons are likely to eventually degrade, inhibiting the opposite transition, from operons back to independent promoters. In total, then, the evolution of SLTS (and operonic systems) are perhaps the best example of recurrent CNE (Lukes et al. 2009), an alternative mechanism to generate increased biological diversity (Stoltzfus 1999; Gray et al. 2010; Doolittle et al. 2011; and see Speijer 2011 for counterarguments).

Massive Loss of U12 Introns

U12 or minor introns are a rare class of introns that are removed by a distinct spliceosomal machinery and characterized by strict extended splice signals. U12 introns are likely to have been present in the last common ancestors of eukaryotes but have been independently reduced in number or completely lost in many lineages (Russell et al. 2005; Alioto 2007; Dávila López et al. 2008; Roy and Irimia 2009b). The dynamics may be governed by a general mutational ratchet (in this case, not associated to CNE): whereas both loss of U12-intron sequences and conversion from U12- to "standard" major U2-spliceosomal introns are routinely observed, and simple mutations causing these changes have been identified in the laboratory, the opposite (U2-to-U12) has never been documented (Burge et al. 1998; Roy and Irimia 2009b).

Case of Recurrent Genome Evolution at the Gene or Gene Family Level

Gene Duplications and Family Expansions

Gene duplication is a frequent phenomenon (Lipinski et al. 2011), which affects a wide variety of gene families and biological processes, suggesting much recurrent gene duplication may be largely stochastic. However, exceptions in which recurrent gene duplication has underpinned parallel phenotypic evolution are also known. One clear example involves duplication of RNAse genes (Zhang 2006). In two lineages of leaf-eating monkeys, a new digestive tract-specific RNAse gene arose by duplication of the same ancestral RNAse and acquired identical amino acid changes altering RNAse activity and resulting in improved leaf digestion. Such cases represent recurrent genomic evolution due to selective environmental pressures acting at on a specific subset of lineages.

Other cases evidence general environmental adaptation by recurrent massive gene family expansion. Some biological functions, such as immunity, chemoreception, and detoxification, require the interaction or the recognition of a vast range of substrates, and, thus, increased molecular diversity of paralogs within the genome could be favored. For instance, cytochrome-P450 genes, which participate in detoxification of various compounds, have undergone pronounced independent expansion in many metazoan lineages (Thomas 2007; Baldwin et al. 2009). A similar situation is found in chordate olfactory receptors, where a correlation with environmental positive-selective pressures is evident (Niimura and Nei 2007; Niimura 2009). On the other hand, other cases of recurrent massive gene family expansion—which are overwhelmingly statistically significant over a random expectation obtained from related gene families—suggest important adaptation of unknown functional significance, raising important questions for further exploration (e.g., EXTK tyrosine kinases, for which dozens of members have independently evolved in several lineages; fig. 1, in contrast to all other related tyrosine kinase families, for which nearly no gene duplications are known in other metazoans lineages, D'Aniello et al. 2008).

Cluster Formation and Assembly of Syntenic Blocks

Pairs or groups of genes may be closely physically linked in different species due to functional reasons. In most cases, this reflects retention of an ancestral association; however,

some instances of repeated evolution of physical linkage between pairs or groups of genes have been described. One set of these involves recurrent evolution of clusters of paralogous genes, presumably by tandem gene duplication and selection against gene translocation. These genomic structures may provide a genetic positive-selective advantage by allowing subtle coding sequence and transcriptional diversification of new gene copies under the control of a shared set of regulatory elements (Tena et al. 2011). Accordingly, many described cases correspond to key developmental genes with complex transcriptional expression patterns (Peterson 2004; Duncan et al. 2008; Irimia et al. 2008; Kuraku et al. 2008; Takatori et al. 2008; Kerner et al. 2009; Negre and Simpson 2009); for example, *Iroquois* genes have independently evolved gene clusters in at least five metazoan lineages (Irimia et al. 2008; Takatori et al. 2008; Kerner et al. 2009), arguing for positive-selective reasons versus stochastic occurrence. More rarely, recurrent linkage of nonparalogous genes may occur, and this association may be favored due to functional advantages (e.g., improved coordination of expression): for instance, for three genes involved in galactose metabolism in two divergent fungal phyla (Slot and Rokas 2010).

Disruption of Highly Conserved Gene Clusters and Other Syntenic Blocks

Ancestral blocks of syntenic genes have been maintained in diverse modern animals, indicating strong selection for their retention in diverse lineages, generally associated with specific developmental programs (e.g., Hox gene clusters; Duboule 2007). However, these associations have been recurrently disrupted in several different animal lineages (Ferrier and Holland 2002; Seo et al. 2004; Pierce et al. 2005; Duboule 2007; Negre and Ruiz 2007). This indicates that these linkages have repeatedly become nonessential, suggesting modification of fundamental animal developmental programs, a potential case of relaxed-selective pressures. Similarly, disruption of ancient associations of phylogenetically unrelated genes, acting as genomic regulatory blocks (Engstrom et al. 2007; Kikuta et al. 2007), have also been reported (e.g., *Iroquois* genes with *Sowah* genes in several lineages; Irimia et al. 2008; Maeso et al. 2012).

Gene Losses

Gene losses constitute an obvious example of nonconstructive mutational ratchet for rather unessential genes. In extreme examples, such as the *GFP* gene family in metazoans and the oxylipin pathway genes in holozoans, the taxonomic distribution implies at least five independent losses (Deheyn et al. 2007; Lee et al. 2008, fig. 1). Alternatively, the loss of the same selective pressure in two lineages due to a common change in lifestyle and/or developmental process (e.g., loss of vision in lightless environments; Protas et al. 2011) may result in dispensability of the same genes and thus in their recurrent loss (environmental relaxed-selection). An example of this is the repeated loss of oxidative phosphorylation complex I genes in anaerobic fungi (Marcet-Houben et al. 2009). In such cases, the loss of one of the genes involved in a particular protein complex or biological pathway could render its interacting partners nonfunctional, further enhancing the loss of the latter. This is exemplified by the absence of all six proteins integrating the fifth adaptor protein (AP-5) complex independently in five different eukaryotic lineages (Hirst et al. 2011).

Genic redundancy, by individual gene duplication or WGD, configures yet another evolutionary scenario for recurrent gene losses (genetic relaxed selection). In these cases, although simple chance is likely to underlie most patterns of gene loss, there are instances in which not all genes seem to be equally prone to retention. For example, some paralogs have been repeatedly lost specifically in different vertebrate lineages, as is the case of *Pdx2* genes in teleosts and tetrapods (Mulley and Holland 2010), *EvxB* in elephant shark and tetrapods (Ravi et al. 2009), *Alx3* in frogs, lizards, and chicken (McGonnell et al. 2011), or globin-E gene (*GbE*) in all major vertebrate lineages but birds (Hoffmann et al. 2011). (It should be noted, however, that although intriguing and suggestive, these patterns of coincidental loss across four/five major vertebrate lineages cannot be statistically significantly different from the null expectation due to the small sample size. Further availability of genomic sequences should overcome this limitation.) More globally, this nonramdom pattern of paralog losses seems to be the rule in yeast (Scannell et al. 2007). Finally, some recurrent losses may reflect positive-selective genetic pressure: for instance, recurrent reduction to a single copy of the same gene families following WGD in plants, fungi, and animals likely reflects strong purifying selection on gene dosage (Paterson et al. 2006).

Cases of Recurrent Evolution of Specific Intragenic Features

Tandem Exon Duplications

Seven to 17% of metazoan genes have tandem exon duplications (Letunic et al. 2002; Gao and Lynch 2009), generally associated with mutually exclusive alternative splicing (Kondrashov and Koonin 2001; Irimia et al. 2008). This alternative processing generates internal redundancy (internal paralogy), which can be exploited to produce functionally divergent transcripts. Although many exon duplications may be (nearly) neutral and occurring by chance, extreme recurrent cases suggest positive-selective forces. A classic example is the *DSCAM* gene, in which exons 6 and 9 have undergone massive, independent expansions in different insect and crustacean lineages (Brites et al. 2008; Lee et al. 2010). Alternative splicing generates many isoforms

of the *DSCAM* gene, which encodes receptors involved in axon guidance, potentially allowing for increased wiring complexity (Schmucker et al. 2000). In the tropomyosin cytoskeletal gene, independent duplication of many different exons has occurred in most bilaterian lineages (Vrhovski et al. 2008; Irimia, Maeso, et al. 2010; Koziol et al. 2011; fig. 1) at a frequency statistically significantly higher than expected even from the highest estimates of intragenic duplications (Gao and Lynch 2009). The explanation appears to lie in the use of alternative promoters to produce two different protein isoforms with radically different cellular functions. Following duplication, each exon copy is "assigned" to one of the two isoforms, reducing pleiotropy and allowing "general positive selection" for optimized function of each protein (Irimia, Maeso, et al. 2010). Finally, another classic example is the parallel evolution of alternative splicing of recurrent tandem exon duplicates in ion channel receptors in flies and mammals (Copley 2004; Fodor and Aldrich 2009).

Gain or Loss of Individual Introns

Intron loss is a relatively common process, especially in some lineages, so the loss of the same intron in a specific gene is likely to occur repeatedly in different lineages simply by chance (Roy and Penny 2006; Roy and Irimia 2008a). However, certain gene features, such as conserved high expression level (Carmel and Koonin 2009), could generate trends toward recurrent intron loss from some genes (a case of general positive selection). Intron gain, on the other hand, is generally thought to be less common, although the extent of parallel gains have been widely debated (e.g., Csurös 2005; Nguyen et al. 2005; Sverdlov et al. 2005), and genome-wide comparisons showed that they may account for up to 8% of the shared intron positions across eukaryotic genes (Carmel et al. 2007). In addition, clear individual cases have been identified (Tarrio et al. 2003; Qiu et al. 2004; Ahmadinejad et al. 2010), even as polymorphisms within populations (Omilian et al. 2008; Li et al. 2009). Nonetheless, despite its lower frequency, parallel intron gain is also likely to occur largely by chance, particularly given than no case of parallel gain in multiple lineages has been described yet. Alternatively, intron gain has long been proposed to be biased toward certain sequences (proto-splice sites; Dibb and Newman 1989), which could impose a general mutational pressure underlying the recurrent patterns.

Recurrent Loss of Gene Parts

Repeated loss of coding sequences of genes may provide parallel changes in protein function or protein–protein interactions (e.g., truncation of C-terminal transactivation domain in *meis/hth* proteins, Irimia et al. 2011; and loss of Snag domains in C_2H_2 zinc fingers, Barrallo-Gimeno and Nieto 2009; Irimia et al. 2010). At the regulatory level, recurrent loss of *cis*-regulatory sequences can have major phenotypic and adaptative consequences with minimal pleiotropic effects (e.g., repeated deletion of a pelvic enhancer in stickleback populations; Chan et al. 2010). In other cases, change in body plans and/or developmental programs may render some regulatory elements unnecessary, even for otherwise deeply conserved sequences (e.g., the only known regulatory element conserved from cnidarians to vertebrates has been lost (or diverged beyond recognition) independently in protostomes, tunicates, and hydra; Royo et al. 2011). Thus, a great variety of causes can be devised for this type of genomic changes, depending on the gene and lineages involved (recurrent-environmental positive-selection, recurrent-environmental and recurrent-genetic relaxed-selection, general mutation, etc.).

Evolution of Coding Sequences

Cases of identical changes in amino acid sequences in different lineages have been extensively studied and represent the paradigmatic example of recurrent molecular phenotypic evolution (Doolittle 1994; Zhang and Kumar 1997; Christin et al. 2010). Parallel amino acid replacements are probably very frequent and happen extensively by chance even at generally highly conserved sites (i.e., "rare amino acid replacements," RGC_CAMs; Irimia et al. 2007; Rogozin et al. 2007a, 2007b, 2008; Roy and Irimia 2008b). However, it has been estimated that homoplastic amino acid substitutions are 2-fold more common than expected under neutral models of protein evolution (Rokas and Carroll 2008). Not surprisingly, then, in addition to the plethora of neutral cases, many studied examples are linked to recurrent environmental positive-selective pressures, with amino acid substitutions conferring adaptative changes to the new environment (e.g., optimal activity at lower pH conditions in the aforementioned RNAses, Zhang 2006; or changes in "hearing genes" in mammals with echolocating systems; Liu et al. 2010; Davies et al. 2011).

The Relationship between Recurrent Genome Evolution and Phenotype

What are the phenotypic effects of this wealth of recurrent genomic changes? It is worth noting that, with regard to the genotype–phenotype map, the study of recurrent genomic changes may be seen as the inverse of the study of recurrent phenotypic changes. The study of recurrent phenotypic evolution is an inherently "top-down" enterprise (fig. 2): study begins with the observation of similar morphological, physiological, or even molecular phenotypes and then investigates whether or not the underlying genetic changes also share similarities (redeployment of the same key developmental genes or similar types of mutations). Recurrent phenotypes may or may not reflect changes in the same pathways, the same genes within those pathways, the same types of changes within those genes (e.g., exon duplication

vs. protein changes), the same specific change (e.g., a specific amino acid change), or the same genome-level change giving rise to the transcript/protein change (e.g., Threonine-to-Serine changes can occur due to substitutions at the first or third codon position). Even if the transcript changes are the same, this could reflect identical or nonidentical changes in the genome (e.g., genomic change vs. RNA editing). In all cases, the organismal phenotypes are equivalent, regardless of the similarity or difference of their genomic bases.

By contrast, study of recurrent genomic evolution is a fundamentally bottom-up pursuit (fig. 2): study begins with an observation of similarity encoded at the genomic level (e.g., independently duplicated exons in tropomyosin genes) and then investigates whether or not these similarities are reflected in resemblance at phenotypic levels (optimization of the same two protein functions). For instance, consider a recurrent intragenic tandem duplication. The duplications may affect the transcriptome or may not (e.g., an intronic duplication may not). Exonic duplications may affect the protein sequence/function/structure or may not (e.g., an exon in a UTR). Protein-affecting changes may or may not affect cellular/organismal phenotype. Fundamentally, then, whereas repeated phenotypic evolution may speak directly of adaptative values, but only rarely (and sometimes indirectly) about the evolutionary mechanisms of genetic change, recurrent genomic evolution directly informs about the genetic changes themselves, although adaptative causes can remain more elusive. The types and extents of phenotypic changes due to recurrent genomic changes—and the similarities of these changes across lineages—remain largely unknown and represent an important set of questions in understanding recurrent evolution.

What Do Recurrent Genomic Features Then Tell Us about Evolution?

Genomic recurrence provides a new perspective on evolutionary processes, informing us in often unexpected ways about commonalities of forces—mutational and/or selectional—acting across different lineages. Cases of genomic recurrence caused by ratchet mutations are fundamental to understanding the evolutionary constraints and canalizations that shape the way in which the "genome-space," as the morphospace, is explored through evolution, underscoring predictability in the overall outcome of neutral mutation, whether or not this will be "constructive" (Stoltzfus 1999; Gray et al. 2010; Doolittle et al. 2011; Speijer 2011). For example, the observation of recurrent emergence of SLTS suggests that the mutational path to a new SLTS system is readily available over long evolutionary times; on the other hand, the lack of reversion from SLTS to non-SLTS presumably indicates general selective forces opposing loss of SLTS, for instance due to loss of the machinery involved in the non–SLTS-dependent expression of the genes subject to SLTS.

Other quasineutral changes that have been repeatedly used as substrate for molecular innovations suggest that certain genomic traits confer evolutionary flexibility, opening new venues that can be explored during evolution. Thus, their mere presence would be indicative of evolutionary potential, allowing specific hypotheses about the occurrence of typically accompanying features (e.g., reorganization of conserved synteny after WGDs or the creation of operons in the presence of SLTS).

In other cases, although cellular/organismal phenotypic consequences of genomic recurrence may not be immediately evident, careful study of genomic patterns can provide straightforward testable hypotheses about phenotypic consequences. For instance, the observation of recurrent evolution of gastrointestinal RNAase paralogs in two leaf-eating monkey lineages made specific predictions that protein sequence changes in the gastrointestinal RNAase gene would enhance digestion, which were later experimentally confirmed (Zhang 2006).

However, it is in the less predictable cases in which the study of recurrent genome evolution arguably reaches the height of its power. For instance, the finding that splicing motifs become highly similar among the remaining introns in nearly intronless species came as a profound surprise (Irimia et al. 2007; Irimia and Roy 2008; Schwartz et al. 2008). This pattern indicates a rule that is at the same time extremely clear and poorly understood: In the context of (or following) nearly complete intron loss, selection for consensus sequences increases on remaining introns. In such cases, the repeatability of the evolutionary outcomes is likely to point at specific ways in how selection acts on these features, illuminating the path for future research.

Concluding Remarks

The diverse instances discussed here represent only a subset of the known cases of repeated evolution at the genome level that have been found largely serendipitously, suggesting that recurrent patterns of genome evolution are widespread. In addition, although recurrent evolution can occur by sheer chance, the above examples provide extensive evidence that genomic recurrence often respond to specific evolutionary forces.

As ancestrally shared features are the result of a common evolutionary history, shared features evolved by recurrent evolution are often the result of common evolutionary forces acting on different lineages. These cases improve our understanding of genome evolution, the causes and the modes, allowing us to make specific predictions about evolutionary outcomes. Unraveling the manifold significance of repeated genomic outputs will necessarily require comprehensive and systematic analyses of recurrent phenomena as well as rigorous statistical testing and greater phylogenetic sampling to assess the dynamics underlying

observed cases of convergence. Given the increasing availability of complete genome sequences, these analyses are increasingly possible, and as with replicates in experimental research, recurrent events will help us to sketch an increasingly focused picture of genome evolution.

Acknowledgments

We thank Eugene V. Koonin and Mike Lynch for helpful comments, and Jordi Paps, Juan Pascual-Anaya, Demián Burguera, Pedro Martinez, Rosario Linacero, Ana Vazquez, and former colleagues from the University of Barcelona for helpful discussions and comments on the manuscript. We would like to apologize to all researchers whose work was not included in this review. I.M. holds a postdoctoral contract funded by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant [268513]11. M.I. is the recipient of a Human Frontiers Science Program Long-Term Fellowship.

Literature Cited

Ahmadinejad N, Dagan T, Gruenheit N, Martin W, Gabaldón T. 2010. Evolution of spliceosomal introns following endosymbiotic gene transfer. BMC Evol Biol. 10:57.

Alioto TS. 2007. U12DB: a database of orthologous U12-type spliceosomal introns. Nucleic Acids Res. 35:D110–D115.

Amores A, et al. 1998. Zebrafish hox clusters and vertebrate genome evolution. Science 282:1711–1714.

Ankarklev J, Jerlström-Hultqvist J, Ringqvist E, Troell K, Svärd SG. 2010. Behind the smile: cell biology and disease mechanisms of Giardia species. Nat Rev Microbiol. 8:413–422.

Arendt J, Reznick D. 2008. Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? Trends Ecol Evol. 23:26–32.

Aury J-M, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature 444: 171–178.

Baldwin W, Marko P, Nelson D. 2009. The cytochrome P450 (CYP) gene superfamily in Daphnia pulex. BMC Genomics 10:169.

Barrallo-Gimeno A, Nieto MA. 2009. Evolutionary history of the Snail/Scratch superfamily. Trends Genet. 25:248–252.

Belle EM, Duret L, Galtier N, Eyre-Walker A. 2004. The decline of isochores in mammals: an assessment of the GC content variation along the mammalian phylogeny. J Mol Evol. 58:653–660.

Bellott DW, et al. 2010. Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. Nature 466: 612–616.

Bennetzen JL. 2002. Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica 115:29–36.

Bergero R, Forrest A, Kamau E, Charlesworth D. 2007. Evolutionary strata on the X chromosomes of the dioecious plant Silene latifolia: evidence from new sex-linked genes. Genetics 175:1945–1954.

Birdsell JA. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol Biol Evol. 19:1181–1197.

Bisbee CA, Baker MA, Wilson AC, Haji-Azimi I, Fischberg M. 1977. Albumin phylogeny for clawed frogs (Xenopus). Science 195:785–787.

Blomme T, et al. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol. 7:R43.

Brites D, et al. 2008. The Dscam homologue of the crustacean Daphnia is diversified by alternative splicing like in insects. Mol Biol Evol. 25: 1429–1439.

Burge CB, Padgett RA, Sharp PA. 1998. Evolutionary fates and origins of U12-type introns. Mol Cell. 2:773–785.

Carmel L, Koonin EV. 2009. A universal nonmonotonic relationship between gene compactness and expression levels in multicellular eukaryotes. Genome Biol Evol. 1:382–390.

Carmel L, Rogozin IB, Wolf YI, Koonin EV. 2007. Patterns of intron gain and conservation in eukaryotic genes. BMC Evol Biol. 7:192.

Chan YF, et al. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science 327:302–305.

Charlesworth D, Mank JE. 2010. The birds and the bees and the flowers and the trees: lessons from genetic mapping of sex determination in plants and animals. Genetics 186:9–31.

Christin PA, Weinreich DM, Besnard G. 2010. Causes and evolutionary significance of genetic convergence. Trends Genet. 26:400–405.

Conway Morris S. 2009. The predictability of evolution: glimpses into a post-Darwinian world. Naturwissenschaften 96:1313–1337.

Copley RR. 2004. Evolutionary convergence of alternative splicing in ion channels. Trends Genet. 20:171–176.

Corradi N, Pombert JF, Farinelli L, Didier ES, Keeling PJ. 2010. The complete sequence of the smallest known nuclear genome from the microsporidian Encephalitozoon intestinalis. Nat Commun. 1:77.

Covello PS, Gray MW. 1993. On the evolution of RNA editing. Trends Genet. 9:265–268.

Csurös M. 2005. Likely scenarios of intron evolution. Third RECOMB Satellite Workshop on Comparative Genomics. Berlin (Germany): Springer LNCS 3678. p. 47–60.

D'Aniello S, et al. 2008. Gene expansion and retention leads to a diverse tyrosine kinase superfamily in amphioxus. Mol Biol Evol. 25: 1841–1854.

Davies KT, Cotton JA, Kirwan JD, Teeling EC, Rossiter SJ. Parallel signatures of sequence evolution among hearing genes in echolocating mammals: an emerging model of genetic convergence. Heredity. Advance Access published December 14, 2011, doi:10.1038/hdy.2011.119

Dávila López M, Rosenblad MA, Samuelsson T. 2008. Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components. Nucleic Acids Res. 36:3001–3010.

Davis JK, Thomas PJ. NISC Comparative Sequencing Program, Thomas JW. 2010. A W-linked palindrome and gene conversion in New World sparrows and blackbirds. Chromosome Res. 18:543–553.

Deheyn DD, et al. 2007. Endogenous green fluorescent protein (GFP) in amphioxus. Biol Bull. 213:95–100.

Dibb NJ, Newman AJ. 1989. Evidence that introns arose at proto-splice sites. EMBO J. 8:2015–2021.

Donoghue PC, Purnell MA. 2005. Genome duplication, extinction and vertebrate evolution. Trends Ecol Evol. 20:312–319.

Doolittle RF. 1994. Convergent evolution: the need to be explicit. Trends Biochem Sci. 19:15–18.

Doolittle WF. 1978. Genes in pieces: were they ever together? Nature 272:581–582.

Doolittle WF. 1998. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet. 14:307–311.

Doolittle WF, Lukes J, Archibald JM, Keeling PJ, Gray MW. 2011. Comment on "Does constructive neutral evolution play an important role in the origin of cellular complexity?" Bioessays 33:427–429.

Douris V, Telford MJ, Averof M. 2010. Evidence for multiple independent origins of trans-splicing in Metazoa. Mol Biol Evol. 27:684–693.
Duboule D. 2007. The rise and fall of Hox gene clusters. Development 134:2549–2560.
Duncan E, Wilson M, Smith J, Dearden P. 2008. Evolutionary origin and genomic organisation of runt-domain containing genes in arthropods. BMC Genomics 9:558.
Duret L. 2006. The GC content of primates and rodents genomes is not at equilibrium: a reply to Antezana. J Mol Evol. 62:803–806.
Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. 4:e1000071.
Duret L, Semon M, Piganeau G, Mouchiroud D, Galtier N. 2002. Vanishing GC-rich isochores in mammalian genomes. Genetics 162:1837–1847.
Edger PP, Pires JC. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Res. 17:699–717.
Eichinger L, et al. 2005. The genome of the social amoeba Dictyostelium discoideum. Nature 435:43–57.
Eisen JA, et al. 2006. Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. PLoS Biol. 4:e286.
Ellegren H. 2011. Sex-chromosome evolution: recent progress and the influence of male and female heterogamety. Nat Rev Genet. 12: 157–166.
Engstrom PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B. 2007. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res. 17:1898–1908.
Evans BJ, Kelley DB, Tinsley RC, Melnick DJ, Cannatella DC. 2004. A mitochondrial DNA phylogeny of African clawed frogs: phylogeography and implications for polyploidy evolution. Mol Phylogenet Evol. 33:197–213.
Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc Natl Acad Sci U S A. 106:5737–5742.
Ferrier DE, Holland PW. 2002. Ciona intestinalis ParaHox genes: evolution of Hox/ParaHox cluster integrity, developmental mode, and temporal colinearity. Mol Phylogenet Evol. 24:412–417.
Fodor AA, Aldrich RW. 2009. Convergent evolution of alternative splices at domain boundaries of the BK channel. Annu Rev Physiol. 71:19–36.
Fraser JA, Heitman J. 2005. Chromosomal sex-determining regions in animals, plants and fungi. Curr Opin Genet Dev. 15:645–651.
Fraser JA, et al. 2004. Convergent evolution of chromosomal sex-determining regions in the animal and fungal kingdoms. PLoS Biol. 2:e384.
Fritz-Laylin LK, et al. 2010. The genome of Naegleria gruberi illuminates early eukaryotic versatility. Cell 140:631–642.
Gallardo MH, Bickham JW, Honeycutt RL, Ojeda RA, Köhler N. 1999. Discovery of tetraploidy in a mammal. Nature 401:341.
Galperin MY, Koonin EV. 2012. Divergence and convergence in enzyme evolution. J Biol Chem. 287:21–28.
Gao X, Lynch M. 2009. Ubiquitous internal gene duplication and intron creation in eukaryotes. Proc Natl Acad Sci U S A. 106: 20818–20823.
Gardner MJ, et al. 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419:498–511.
Ghedin E, et al. 2007. Draft genome of the filarial nematode parasite Brugia malayi. Science 317:1756–1760.
Gray MW, Lukes J, Archibald JM, Keeling PJ, Doolittle WF. 2010. Irremediable complexity? Science 330:920–921.
Gregory TR, et al. 2007. Eukaryotic genome size databases. Nucleic Acids Res 35:D332–D338.
Gruber KF, Voss RS, Jansa SA. 2007. Base-compositional heterogeneity in the RAG1 locus among didelphid marsupials: implications for phylogenetic inference and the evolution of GC content. Syst Biol. 56:83–96.
Haag ES. 2005. The evolution of nematode sex determination: C. elegans as a reference point for comparative biology. WormBook: The C. elegans Research Community.
Hershberg R, Petrov DA. 2009. General rules for optimal codon choice. PLoS Genet. 5:e1000556.
Hirst J, et al. 2011. The fifth adaptor protein complex. PLoS Biol. 9: e1001170.
Hoffmann FG, Opazo JC, Storz JF. 2011. Differential loss and retention of cytoglobin, myoglobin, and globin-E during the radiation of vertebrates. Genome Biol Evol. 3:588–600.
Hurley IA, et al. 2007. A new time-scale for ray-finned fish evolution. Proc R Soc B Biol Sci. 274:489–498.
Irimia M, Maeso I, Garcia-Fernandez J. 2008. Convergent evolution of clustering of Iroquois homeobox genes across metazoans. Mol Biol Evol. 25:1521–1525.
Irimia M, Maeso I, Gunning PW, Garcia-Fernandez J, Roy SW. 2010. Internal and external paralogy in the evolution of tropomyosin genes in metazoans. Mol Biol Evol. 27:1504–1517.
Irimia M, Maeso I, Penny D, Garcia-Fernandez J, Roy SW. 2007. Rare coding sequence changes are consistent with Ecdysozoa, not Coelomata. Mol Biol Evol. 24:1604–1607.
Irimia M, Penny D, Roy SW. 2007. Co-evolution of genomic intron number and splice sites. Trends Genet. 23:321–325.
Irimia M, Roy SW. 2008. Evolutionary convergence on highly-conserved 3′ intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. PLoS Genet. 4:e1000148.
Irimia M, et al. 2008. Widespread evolutionary conservation of alternatively spliced exons in Caenorhabditis. Mol Biol Evol. 25: 375–382.
Irimia M, et al. 2009. Complex selection on 5′ splice sites in intron-rich organisms. Genome Res. 19:2021–2027.
Irimia M, et al. 2010. Conserved developmental expression of Fezf in chordates and Drosophila and the origin of the Zona Limitans Intrathalamica (ZLI) brain organizer. EvoDevo 1:7.
Irimia M, et al. 2011. Contrasting 5′ and 3′ evolutionary histories and frequent evolutionary convergence in Meis/hth gene structures. Genome Biol Evol. 3:551–564.
Kaiser VB, Bachtrog D. 2010. Evolution of sex chromosomes in insects. Annu Rev Genet. 44:91–112.
Kennedy B, Sabara H, Haydon D, Husband B. 2006. Pollinator-mediated assortative mating in mixed ploidy populations of Chamerion angustifolium (Onagraceae). Oecologia 150:398–408.
Kerner P, Ikmi A, Coen D, Vervoort M. 2009. Evolutionary history of the iroquois/Irx genes in metazoans. BMC Evol Biol. 9:74.
Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica 115:49–63.
Kikuta H, et al. 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res. 17:545–555.
Kondrashov FA, Koonin EV. 2001. Origin of alternative splicing by tandem exon duplication. Hum Mol Genet. 10:2661–2669.
Koziol U, et al. 2011. Developmental expression of high molecular weight tropomyosin isoforms in Mesocestoides corti. Mol Biochem Parasitol. 175:181–191.
Kraut J. 1977. Serine proteases: structure and mechanism of catalysis. Annu Rev Biochem. 46:331–358.

Kuraku S, et al. 2008. Noncanonical role of Hox14 revealed by its expression patterns in lamprey and shark. Proc Natl Acad Sci U S A. 105:6679–6683.
Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.
Lane CE, et al. 2007. Nucleomorph genome of Hemiselmis andersenii reveals complete intron loss and compaction as a driver of protein structure and function. Proc Natl Acad Sci U S A. 104:19908–19913.
Lang D, et al. 2010. Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. Genome Biol Evol. 2:488–503.
Lee C, Kim N, Roy M, Graveley BR. 2010. Massive expansions of Dscam splicing diversity via staggered homologous recombination during arthropod evolution. RNA 16:91–105.
Lee D-S, Nioche P, Hamberg M, Raman CS. 2008. Structural insights into the evolutionary paths of oxylipin biosynthetic enzymes. Nature 455: 363–368.
Letunic I, Copley RR, Bork P. 2002. Common exon duplication in animals and its role in alternative splicing. Hum Mol Genet. 11:1561–1567.
Li W, Tucker AE, Sung W, Thomas WK, Lynch M. 2009. Extensive, recent intron gains in Daphnia populations. Science 326:1260–1262.
Lipinski KJ, et al. 2011. High spontaneous rate of gene duplication in Caenorhabditis elegans. Curr Biol. 21:306–310.
Liu Y, et al. 2010. Convergent sequence evolution between echolocating bats and dolphins. Curr Biol. 20:R53–R54.
Lukes J, Leander BS, Keeling PJ. 2009. Cascades of convergent evolution: the corresponding evolutionary histories of euglenozoans and dinoflagellates. Proc Natl Acad Sci U.S.A. 106:9963–9970.
Lynch M. 2006. The origins of eukaryotic gene structure. Mol Biol Evol. 23:450–468.
Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.
Lynch M. 2011. Statistical inference on the mechanisms of genome evolution. PLoS Genet. 7:e1001389.
Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302:1401–1404.
Ma L-J, et al. 2009. Genomic analysis of the basal lineage fungus *Rhizopus oryzae* reveals a whole-genome duplication. PLoS Genet. 5:e1000549.
Maeso I, et al. 2012. An ancient genomic regulatory block conserved across bilaterians and its dismantling in tetrapods by retrogene replacement. Genome Res. 22:642–655.
Marcet-Houben M, Marceddu G, Gabaldon T. 2009. Phylogenomics of the oxidative phosphorylation in fungi reveals extensive gene duplication followed by functional divergence. BMC Evol Biol. 9:295.
Martens C, Van de Peer Y. 2010. The hidden duplication past of the plant pathogen Phytophthora and its consequences for infection. BMC Genomics 11:353.
Matsuzaki M, et al. 2004. Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. Nature 428: 653–657.
McGonnell IM, et al. 2011. Evolution of the Alx homeobox gene family: parallel retention and independent loss of the vertebrate Alx3 gene. Evol Dev. 13:343–351.
McLean CY, et al. 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature 471:216–219.
Merchant SS, et al. 2007. The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science 318: 245–250.
Morrison HG, et al. 2007. Genomic minimalism in the early diverging intestinal parasite Giardia lamblia. Science 317:1921–1926.
Mulley JF, Holland PWH. 2010. Parallel retention of Pdx2 genes in cartilaginous fish and coelacanths. Mol Biol Evol. 27:2386–2391.
Navajas-Pérez R, et al. 2005. The evolution of reproductive systems and sex-determining mechanisms within rumex (polygonaceae) inferred from nuclear and chloroplastidial sequence data. Mol Biol Evol. 22:1929–1939.
Negre B, Ruiz A. 2007. HOM-C evolution in Drosophila: is there a need for Hox gene clustering? Trends Genet. 23:55–59.
Negre B, Simpson P. 2009. Evolution of the achaete-scute complex in insects: convergent duplication of proneural genes. Trends Genet. 25:147–152.
Nguyen H, Yoshihama M, Kenmochi N. 2005. New maximum likelihood estimators for eukaryotic intron evolution. PLoS Comput Biol. 1:e79.
Niimura Y. 2009. On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. Genome Biol Evol. 1:34–44.
Niimura Y, Nei M. 2007. Extensive gains and losses of olfactory receptor genes in mammalian evolution. PLoS One 2:e708.
Omilian AR, Scofield DG, Lynch M. 2008. Intron presence-absence polymorphisms in Daphnia. Mol Biol Evol. 25:2129–2139.
Opperman CH, et al. 2008. Sequence and genetic map of Meloidogyne hapla: a compact nematode genome for plant parasitism. Proc Natl Acad Sci U.S.A. 105:14802–14807.
Paterson AH, et al. 2006. Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon. Trends Genet. 22:597–602.
Peterson KJ. 2004. Isolation of Hox and Parahox genes in the hemichordate Ptychodera flava and the evolution of deuterostome Hox genes. Mol Phylogenet Evol. 31:1208–1215.
Petrov DA. 2002a. DNA loss and evolution of genome size in Drosophila. Genetica 115:81–91.
Petrov DA. 2002b. Mutational equilibrium model of genome size evolution. Theor Popul Biol. 61:531–544.
Piegu B, et al. 2006. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Res. 16:1262–1269.
Pierce RJ, et al. 2005. Evidence for a dispersed Hox gene cluster in the platyhelminth parasite Schistosoma mansoni. Mol Biol Evol. 22:2491–2503.
Protas ME, Trontelj P, Patel NH. 2011. Genetic basis of eye and pigment loss in the cave crustacean, Asellus aquaticus. Proc Natl Acad Sci U S A. 108:5702–5707.
Putnam NH, et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86–94.
Qiu W, Schisler N, Stoltzfus A. 2004. The evolutionary gain of spliceosomal introns: sequence and phase preferences. Mol Biol Evol. 21:1252–1263.
Ravi V, et al. 2009. Elephant shark (Callorhinchus milii) provides insights into the evolution of Hox gene clusters in gnathostomes. Proc Natl Acad Sci U S A. 106:16327–16332.
Rho M, et al. 2009. Independent mammalian genome contractions following the KT boundary. Genome Biol Evol. 1:2–12.
Rogozin IB, Thomson K, Csürös M, Carmel L, Koonin EV. 2008. Homoplasy in genome-wide analysis of rare amino acid replacements: the molecular-evolutionary basis for Vavilov's law of homologous series. Biol Direct. 3:7.
Rogozin IB, Wolf YI, Carmel L, Koonin EV. 2007a. Analysis of rare amino acid replacements supports the Coelomata clade. Mol Biol Evol. 24:2594–2597.

Rogozin IB, Wolf YI, Carmel L, Koonin EV. 2007b. Ecdysozoan clade rejected by genome-wide analysis of rare amino acid replacements. Mol Biol Evol. 24:1080–1090.
Rokas A, Carroll SB. 2008. Frequent and widespread parallel evolution of protein sequences. Mol Biol Evol. 25:1943–1953.
Romiguier J, Ranwez V, Douzery EJ, Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. Genome Res. 20:1001–1009.
Roy SW, Irimia M. 2008a. Rare genomic characters do not support Coelomata: intron loss/gain. Mol Biol Evol. 25:620–623.
Roy SW, Irimia M. 2008b. Rare genomic characters do not support Coelomata: RGC_CAMs. J Mol Evol. 66:308–315.
Roy SW, Irimia M. 2009a. Mystery of intron gain: new data and new models. Trends Genet. 25:67–73.
Roy SW, Irimia M. 2009b. Splicing in the eukaryotic ancestor: form, function and dysfunction. Trends Ecol Evol. 24:447–455.
Roy SW, Penny D. 2006. Smoke without fire: most reported cases of intron gain in nematodes instead reflect intron losses. Mol Biol Evol. 23:2259–2262.
Roy SW, Penny D. 2007. Widespread intron loss suggests retrotransposon activity in ancient apicomplexans. Mol Biol Evol. 24:1926–1933.
Royo JL, et al. 2011. Transphyletic conservation of developmental regulatory state in animal evolution. Proc Natl Acad Sci U S A. 108:14186–14191.
Russell A, Shutt T, Watkins R, Gray M. 2005. An ancient spliceosomal intron in the ribosomal protein L7a gene (Rpl7a) of Giardia lamblia. BMC Evol Biol. 5:45.
Scannell DR, et al. 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proc Natl Acad Sci U S A. 104:8397–8402.
Schmucker D, et al. 2000. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell 101:671–684.
Schwartz S, et al. 2008. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. Genome Res. 18:88–103.
Scotland RW. 2011. What is parallelism? Evol Dev. 13:214–227.
Seo HC, et al. 2004. Hox cluster disintegration with persistent anteroposterior order of expression in Oikopleura dioica. Nature 431:67–71.
Slamovits CH, Keeling PJ. 2009. Evolution of ultrasmall spliceosomal introns in highly reduced nuclear genomes. Mol Biol Evol. 26:1699–1705.
Slot JC, Rokas A. 2010. Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi. Proc Natl Acad Sci U S A. 107:10136–10141.
Speijer D. 2011. Does constructive neutral evolution play an important role in the origin of cellular complexity? Making sense of the origins and uses of biological complexity. Bioessays 33:344–349.
Stoltzfus A. 1999. On the possibility of constructive neutral evolution. J Mol Evol. 49:169–181.
Sverdlov A, Rogozin I, Babenko V, Koonin E. 2005. Conservation versus parallel gains in intron evolution. Nucleic Acids Res. 33:1741–1748.
Taft RJ, Pheasant M, Mattick JS. 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays 29: 288–299.
Takatori N, et al. 2008. Comprehensive survey and classification of homeobox genes in the genome of amphioxus, Branchiostoma floridae. Dev Genes Evol. 218:579–590.
Tarrio R, Rodríguez-Trelles F, Ayala FJ. 2003. A new Drosophila spliceosomal intron position is common in plants. Proc Natl Acad Sci U S A. 100:6580–6583.
Tena JJ, et al. 2011. An evolutionarily conserved three-dimensional structure in the vertebrate Irx clusters facilitates enhancer sharing and co-regulation. Nat Commun. 2:310.
Thomas JH. 2007. Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. PLoS Genet. 3:e67.
Thompson JN, Merg KF. 2008. Evolution of polyploidy and the diversification of plant-pollinator interactions. Ecology 89: 2197–2206.
Ungerer MC, Strakosh SC, Zhen Y. 2006. Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. Curr Biol. 16:R872–R873.
Vanacova S, Yan W, Carlton JM, Johnson PJ. 2005. Spliceosomal introns in the deep-branching eukaryote Trichomonas vaginalis. Proc Natl Acad Sci U S A. 102:4430–4435.
Venkatesh B, Gilligan P, Brenner S. 2000. Fugu: a compact vertebrate reference genome. FEBS Lett. 476:3–7.
Venter JC, et al. 2001. The sequence of the human genome. Science 291:1304–1351.
Vogel C, Chothia C. 2006. Protein family expansions and biological complexity. PLoS Comput Biol. 2:e48.
Vrhovski B, Thézé N, Thiébaud P. 2008. Structure and evolution of tropomyosin genes. Adv Exp Med Biol. 644:6–26.
Wang H, et al. 2005. The origin of the naked grains of maize. Nature 436:714–719.
Whitney KD, Garland T. 2010. Did genetic drift drive increases in genome complexity? PLoS Genet. 6:e1001080.
Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713.
Yampolsky LY, Stoltzfus A. 2001. Bias in the introduction of variation as an orienting factor in evolution. Evol Dev. 3:73–83.
Zhang G, Cohn MJ. 2008. Genome duplication and the origin of the vertebrate skeleton. Curr Opin Genet Dev. 18:387–393.
Zhang J. 2006. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. Nat Genet. 38:819–823.
Zhang J, Kumar S. 1997. Detection of convergent and parallel evolution at the amino acid sequence level. Mol Biol Evol. 14:527–536.
Zhang J, et al. 2010. Loss of fish actinotrichia proteins and the fin-to-limb transition. Nature 466:234–237.

Associate editor: Purificación López-García