Cell. Mol. Life Sci. 64 (2007) 542 – 554
1420-682X/07/050542-13
DOI 10.1007/s00018-006-6453-4

Cellular and Molecular Life Sciences

Review

Current topics in genome evolution: Molecular mechanisms of new gene formation

D. V. Babushok[a], E. M. Ostertag[a, b] and H. H. Kazazian, Jr[a, *]

[a] Department of Genetics, University of Pennsylvania School of Medicine, 475 Clinical Research Building, 415 Curie Blvd, Philadelphia, Pennsylvania 19104 – 6145 (USA), Fax: + 1 215 573 7760,
e-mail: kazazian@mail.med.upenn.edu
[b] Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA (USA)

Received 18 October 2006; received after revision 18 November 2006; accepted 28 November 2006
Online First 27 December 2006

Abstract. Comparative genome analyses reveal that most functional domains of human genes have homologs in widely divergent species. These shared functional domains, however, are differentially shuffled among evolutionary lineages to produce an increasing number of domain architectures. Combined with duplication and adaptive evolution, domain shuffling is responsible for the great phenotypic complexity of higher eukaryotes. Although the domain-shuffling hypothesis is generally accepted, determining the molecular mechanisms that lead to domain shuffling and novel gene creation has been challenging, as sequence features accompanying the formation of known genes have been obscured by accumulated mutations. The growing availability of genome sequences and EST databases allows us to study the characteristics of newly emerged genes. Here we review recent genome-wide DNA and EST analyses, and discuss the three major molecular mechanisms of gene formation: (1) atypical spicing, both within and between genes, followed by adaptation, (2) tandem and interspersed segmental duplications, and (3) retrotransposition events.

Keywords. New gene formation, exon-shuffling, alternative splicing, gene duplication, retrotransposition.

Introduction

The development and optimization of sequencing technology and alignment algorithms have led to the explosion of high-quality whole genome sequence data over the last decade. By the fall of 2006, whole genome sequences of 387 bacteria, 29 archaea, and 44 eukaryotes, including 7 vertebrates (5 mammals [1–6], 1 bird [7] and 1 fish [8]) have been published, and assembly-stage sequences for dozens more are publicly available [9]. The abundance and depth of the nucleotide and protein data, easily accessible alongside critical analysis information such as phylogenetic conservation, domain structure, and repeat content in public data banks worldwide [10–13], has allowed rapid progress in the biomedical fields. Identification of candidate disease genes, mutation screening, novel gene function prediction, and selection of appropriate model organisms are but a few of the applications made routine by sequence availability [1].

Importantly, in addition to constituting an irreplaceable research tool, the availability of vertebrate genome sequences has permitted, for the first time, detailed comparison of the genomes of closely related as well as divergent vertebrate species. Perhaps the

* Corresponding author.

most critical problems to tackle in cross-species genome comparisons are: (1) the discovery and annotation of functional genomic regions, and (2) elucidation of the commonalities and differences that make for the identity of each species. Somewhat unexpectedly, the complexity of eukaryotic genes, compounded by the enormous repetitive content of mammalian genomes, has frustrated the best efforts to obtain a clear 'periodic table'-like compilation of mammalian genes so far [1, 14]. Experimental validation, combined with continued advances in bioinformatics and completion of a larger set of genome sequences are certain to clarify and expand our current genome annotation. Nevertheless, painstaking analyses of the completed genomes have already led to key insights into the nature of mammalian genomes, their complexity and composition, and the dynamic processes that power evolution.

In this review, we discuss the extensive commonalities in the protein-coding sequence among divergent vertebrates, with a particular focus on the expansion of gene families and molecular mechanisms of novel gene formation.

Protein-coding genes are conserved among vertebrates

Reciprocal nucleotide alignments of sequenced genomes have narrowed down the core genomic regions inherited from the common vertebrate ancestor. As expected, the amount of sequence that can be aligned to the human genome generally decreases with increasing time since divergence from humans. While best reciprocal nucleotide alignments with the human genome cover the vast majority of assembled chimpanzee genome [5], the human-alignable fraction decreases to ~40% in mouse and rat [2, 3], ~60% in dog, 34% in marsupials and 14% in platypus genomes [15, 16]. Despite the fact that rodents are evolutionarily closer to humans than dog, significant deletions in the rodent lineage resulted in a lower than expected fraction of human-alignable sequence [3, 15].

It is useful to consider these shared regions as a continually distilled subset of the core vertebrate genome, which, at its limit, should contain the most essential elements of an ancestral vertebrate. Thus, the human-alignable regions in the chimpanzee contain as much as 40% of shared repetitive content and 60% non-annotated sequences in addition to the full complement of protein-coding regions. With each additional comparison, the two former categories decrease, comprising roughly 6% and 32% in rodents, respectively, while their protein-coding fraction remains largely preserved [15, 16]. Consistent with this, the human-aligned regions for the furthest diverged vertebrates studied (birds and fish) are virtually limited to the protein-coding sequences; importantly, the vast majority (98%) of human coding genes remain conserved in chicken through 310 million years (Myr) of separate evolution, and over 68% are conserved in 450 Myr-diverged pufferfish [7].

The selective constraint that underlies the gradual enrichment of protein-coding sequences is most evident in their low divergence between the diverse vertebrates. In contrast, both the repetitive and non-annotated regions undergo generally neutral evolution and are greatly divergent [15]. Notably, as comparisons are made across genomes separated by large evolutionary distances, nonlinear evolution along the branches of the evolutionary tree becomes apparent: the human-alignable sequence from the more evolutionarily distant species is not fully contained within the alignable sequence of more closely related species [7, 16], indicating lineage-specific genome expansions and sequence losses [7].

Great gene complexity created from a limited repertoire of inherited genes

The repertoire of mammalian protein-coding genes is strikingly similar, suggesting that the enormous interspecies phenotypic variation is caused by elaboration on the existing gene structures rather than by *de novo* invention of genes. The closely related chimpanzee genome contains no genes lacking a human homolog, and the typical chimpanzee gene differs from its human counterpart by only two amino acids; 27% of chimpanzee-human pairs are identical [5]. Similarly, over 99% of mouse genes have a human homolog, and the remaining 1% of genes have homology in other mammals [3]. From the available data, only 10 rodent genes without homologs in completed genomes have been identified [2, 3], and their number may decrease as additional mammalian genomes become available. Importantly, while the majority of vertebrate genes arose from a common ancestor, lineage-specific gene losses and gene family expansions produced a tremendous cross-species variability in the relative contributions of different gene families to the species' total coding DNA [1–7]. In fact, only one quarter to one third of the estimated 20 000–30 000 human genes [1] are postulated to make up the conserved, orthologous vertebrate gene core, a hypothesis supported by only 7606 strict 1 : 1 : 1 orthologs in a human-chicken-fugu comparison [7]. The majority of human genes share more complex 1 : n, n : 1, or n : n relationships with their homologs in other species.

Table 1. Comparative analysis of gene and repeat content of the human, fruitfly, worm and yeast genomes.

	Genome size (Mb)[1]	Total no. of genes/ % duplicated genes[2]	Fraction of proteome with human homologs	No. of domain types[3]	Distinct domain architectures (two or more domains)[4]	% of interspersed repeats[5]
S. cerevisiae	12	6680/30%	46%	973	470	3.1%
C. elegans	100	20 060/49%	43%	1183	1248	6.5%
D. melanogaster	133	14 039/41%	61%	1218	1702	3.1%
H. sapiens	3,093	23 224/38%	n/a	1865	3433	46.4%

[1] Golden Path Lengths for yeast, worm, fruitfly and human genomes were obtained from the following Assembly/Genebuild/Ensembl versions: SGD 1 (Nov 2005)/SGD (Nov 2005)/41.1d, WS 160 (July 2006)/Wormbase (May 2006)/41.16, BDGP 4.3 (July 2005)/ FlyBase (Mar 2006)/41.43, and NCBI 36 (Oct 2005)/Ensembl(Aug 2006)/41.36c, respectively.
[2] Total number of genes is comprised of Known Genes and Novel Genes as annotated by Ensembl genome browser (see [1] above). Percent of duplicated genes: [1, 74, 130].
[3] [1].
[4] [20].
[5] [1, 3, 131].

Several notable examples of lineage-specific gene expansions illuminate the physiological adaptations of their host species. For example, the emergence of the mammal-specific vomeronasal organ receptor family correlates with the development of the vomeronasal organ in mammals [7]. Similarly, mammalian expansion of genes involved in olfaction, reproduction, and immunity are thought to mediate significant physiological changes in mammals in these areas [1–3, 5]. One of the most intriguing differences arising from comparisons of hominoid and murid genomes is the accelerated evolution in hominoids of several gene families with roles in the cellular transport of ions and metabolites, synaptic transmission and second messenger-mediated signaling [5]. While genomes of additional species will undoubtedly clarify whether the observed differences are truly hominoid specific [4], it is tempting to speculate on the role these changes played in the development of higher brain function in primates [5].

Increasing organismal complexity was achieved by duplications and progressively complex gene architectures

In The Origin of Species [17], Darwin wrote: "it is quite probable that natural selection, during a long-continued course of modification, should have seized on a certain number of the primordially similar elements, many times repeated, and have adapted them to the most diverse purposes. And as the whole amount of modification will have been effected by slight successive steps, we need not wonder at discovering in such parts or organs, a certain degree of fundamental resemblance, retained by the strong principle of inheritance".

While Darwin's observations describe the phenotypic consequences of actions of multiple genes (*e.g.*, multiplicity of vertebrae, or similarities in whirls of leaves), similar reasoning has been employed over the last century to explain the huge leap in organism complexity arising from a small number of ancestral genes (reviewed in [18]).

In 1990, Dorit and colleagues [19] postulated, based on protein homology analysis, that the whole human gene complement could arise from a limited set of 1000–7000 exon subunits. This estimate was further refined using the whole genome sequence to the 1865 distinct (often multi-exon) domain families in human; 1218 in fruitfly, 1183 in worm, and 973 in yeast (Table 1) [20]. Importantly, almost a complete set of human gene domains is common to one or more lower eukaryotes, and, conversely, a striking 61%, 43%, and 46% of the fly, worm, and yeast proteomes, respectively, is retained in humans [1]. This limited set of shared domains, however, is arranged in an increasing number of combinations and multiplicities, creating 3433 distinct domain architectures of two or more domains in human, compared to the 1702, 1248, and 470 for fly, worm and yeast, respectively [20].

Combining existing domains in novel gene architectures, also known as exon or domain shuffling, has been estimated to have involved up to 20% of eukaryotic exons [21]. The signs of ancient domain shuffling can be detected in current-day eukaryotes as the predominance of intron boundaries in linker regions connecting the domains [22], the symmetrical intron phases at domain boundaries [23], and the correlation of age-prevalent symmetrical intron phases and the age of their protein domains [24]. Moreover, there are several hundred cases of modular proteins thought to have arisen by exon shuffling; most notable examples include blood coagulation

factors (*e.g.*, factors V, VIII, XIII b, protein S) and extracellular matrix constituents (*e.g.*, laminins, collagens, fibrillin, fibronectin) [25, 26]. Thus, extensive domain shuffling resulted in a great boost in complexity of domain architectures, and, combined with differential gene family expansions and their subsequent adaptive evolution, it provides the best working explanation for the N-value paradox [27] – an apparent disconnection between a greatly increased phenotypic complexity of higher eukaryotes and their relatively small number of seemingly derivative genes [1].

Novel genes are formed by atypical splicing of existing genes, duplication, and retrotransposition

Although the domain-shuffling hypothesis is generally accepted, determining the molecular mechanisms that led to domain shuffling and novel gene creation has been challenging, as sequence features accompanying the formation of known genes have been obscured or erased by accumulated mutations [28]. These limitations were partly overcome by the growing availability of genome and transcript sequence, allowing investigators to identify and evaluate the characteristics of young, recently emerged genes. Dozens of individual novel gene studies, together with recent genome-wide surveys of pseudogenes and EST databases point to the following three general mechanisms that alone, or in combination, give rise to new genes: (1) atypical splicing of existing genes, selected over time, (2) DNA duplication, and (3) retrotransposition. A fourth mechanism, acquisition of genes by horizontal transmission, is a well-known force in prokaryote evolution, but is not known to have contributed to genome evolution in vertebrates [28–33], and will not be discussed further here.

Atypical splicing of existing genes

Analyses of EST databases and exon-junction microarrays have revealed that as many as 42–74% of all human genes have at least two splice variants [1, 34, 35], an increase from the estimated 40% and 22% of alternatively spliced genes in fly and worm, respectively [1, 36]; in yeast only a few genes contain introns and are spliced. The repertoire of alternative splicing events (reviewed in [37]) (see Fig. 1) includes omission of exons (Fig. 1b), incorporation of alternative exons (Fig. 1c), alternative transcriptional starts and first exon (Fig. 1d), premature termination with alternative polyadenylation (pA) signal (Fig. 1e), and subtle modifications by alternative splice acceptor or splice donor sites within the framework of the existing exons (Fig. 1f, g).

These events represent a tremendous addition to the coding capacity and functional interactions within the human proteome [38]. While some alternative splicing events are thought to result from stochastic variation in spliceosome binding [37], there is evidence for differential distribution of alternatively spliced genes among different tissues, with the highest prevalence reported in the functionally complex brain, testis, and liver [35, 39]. Furthermore, several cases where alternative splicing is tightly regulated are known, including α-tropomyosin [40], troponin T [41], and PTCH1 oncogene [42].

In addition to the classic form of alternative splicing, which affects a transcript of a single gene, there is now mounting evidence of intergenic splicing events that arise within a read-through chimeric transcript of two different genes located in tandem [43–58]. The majority of such intergenic splicing events, also known as transcription-induced chimeras or TICs, involve tandem transcripts located less than 8.5 kb apart (although ~5% involve genes separated by over 50 kb), and represent as diverse a group of splicing events as traditional alternative splicing (see Fig. 2) [45, 46]. Intergenic splicing is estimated to affect at least 5% of tandemly located genes [46], and up to 2% of all human genes [45]. Careful comparisons of each TIC event in a range of species will be necessary to determine whether it represents a novel transcriptional fusion of two genes, or whether, in fact, it is a remnant of a larger ancestral gene, which has undergone alternative splicing-induced gene fission to form two genes, with the second half of the original gene acquiring independent transcription from a downstream promoter.

Notably, the combination of single-gene alternative splicing and intergenic splicing represents the most common mechanism of novel protein creation. While these creative splicing processes do not generally alter the total gene number (except in extreme cases when TICs evolve into outright gene fusions, or alternative splicing leads to gene fission), they are frequent, ubiquitous and create innovative gene products at low cost to the host. In fact, inclusion of intronic sequences as alternative exons is the only known context for *de novo* exon origination; approximately 2300–2700 novel exons in rodent transcriptome are thought to have been formed by this mechanism [59, 60]. The transposable element sequences found in up to ~4% of human transcripts (typically in minor and/or untranslated RNA isoforms) and in 0.1% of functional human proteins [61–65] had also likely originated as alternatively spliced exons. In agreement with this concept, multiple functional splice sites in the com-

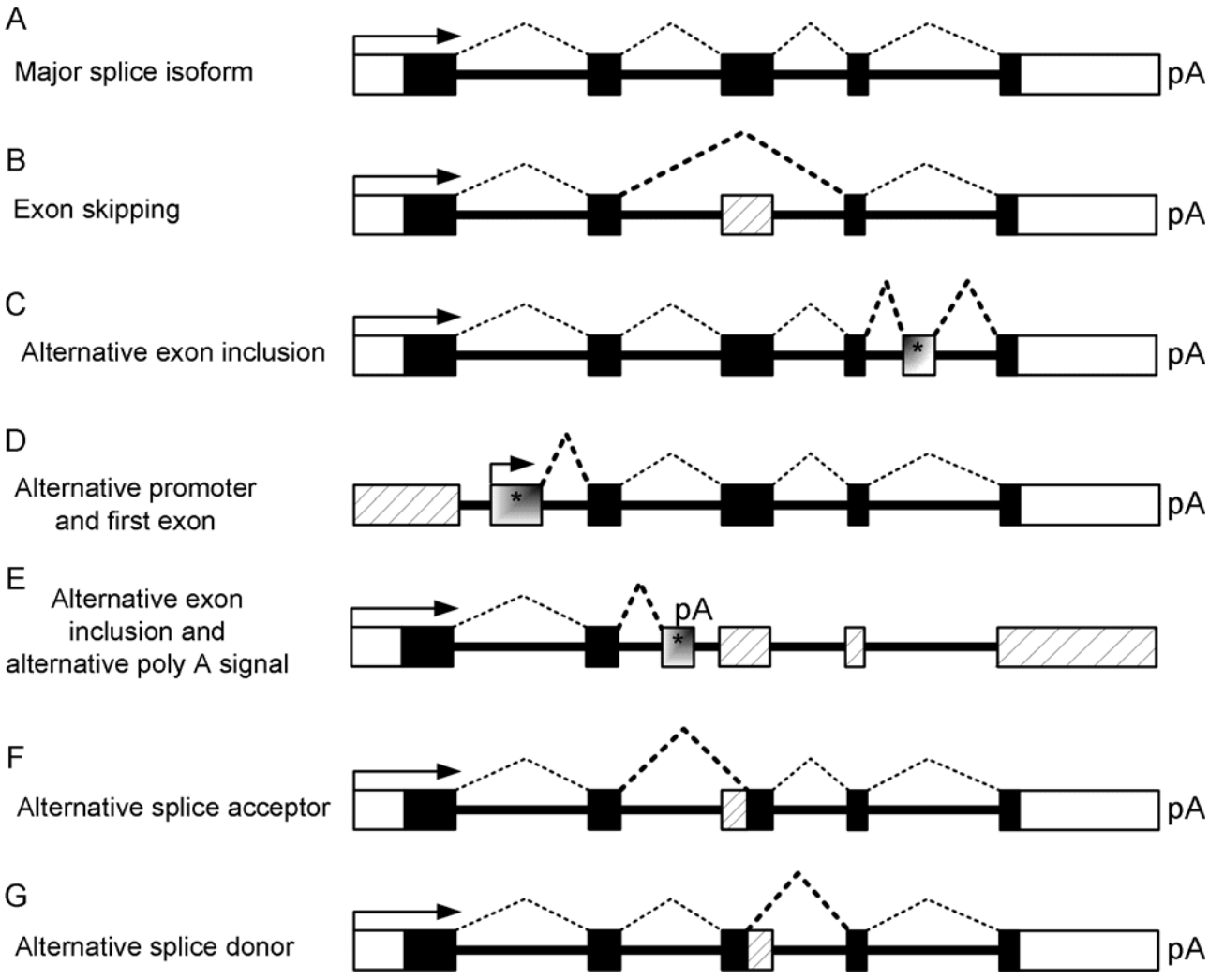

Figure 1. Alternative splicing events within a single gene (reviewed in [37]) include (*b*) skipping of a canonical exon, (*c*) inclusion of an alternative exon, (*d*) transcription from an alternative promoter and alternative first exon, (*e*) inclusion of an alternative terminal exon, and use of (*f*) an alternative splice acceptor or (*g*) splice donor sites within an existing exon. Exons are shown as rectangles, with white and black shading indicating untranslated and coding regions, respectively. Skipped and alternatively included exons are indicated by stripes and gray shading (*), respectively. Bent arrow, transcription start site; pA, polyadenylation signal; dotted lines between exons, splicing.

mon Alu and L1 repeats have been reported [61, 63, 66].

Importantly, ~35% of alternatively spliced transcripts [67] and 56–64% of TICs [45, 46] cause a frame-shift and a premature stop codon, and are expected to undergo nonsense mediated decay (NMD) [68]. Nevertheless, rapid evolution of new exons occasionally creates functional protein-coding regions that may offer a selective advantage to the host; such alternatively spliced transcripts could then evolve into predominant splice variants, or could specialize in tissue- or developmental-specific functions [63]. In contrast to the other two mechanisms of new gene formation – DNA duplication and retrotransposition (discussed below), which do not change the existing gene complement upon new gene creation – an atypical splice product may become a major splice isoform, replacing an existing predominant gene product, which may then become a minor isoform and be lost overtime.

DNA duplication

The requirement for gene duplication to explain the gene number expansion in the evolution of eukaryotes was postulated well before the last decade (reviewed in [18]), when extensive amounts of duplicated genes became apparent in genomes of organisms as diverse as mycoplasma and humans (Table 1). Duplications of whole genomes has been reported for a number of species, including *Saccharomyces cerevisiae* [69], *Ara-*

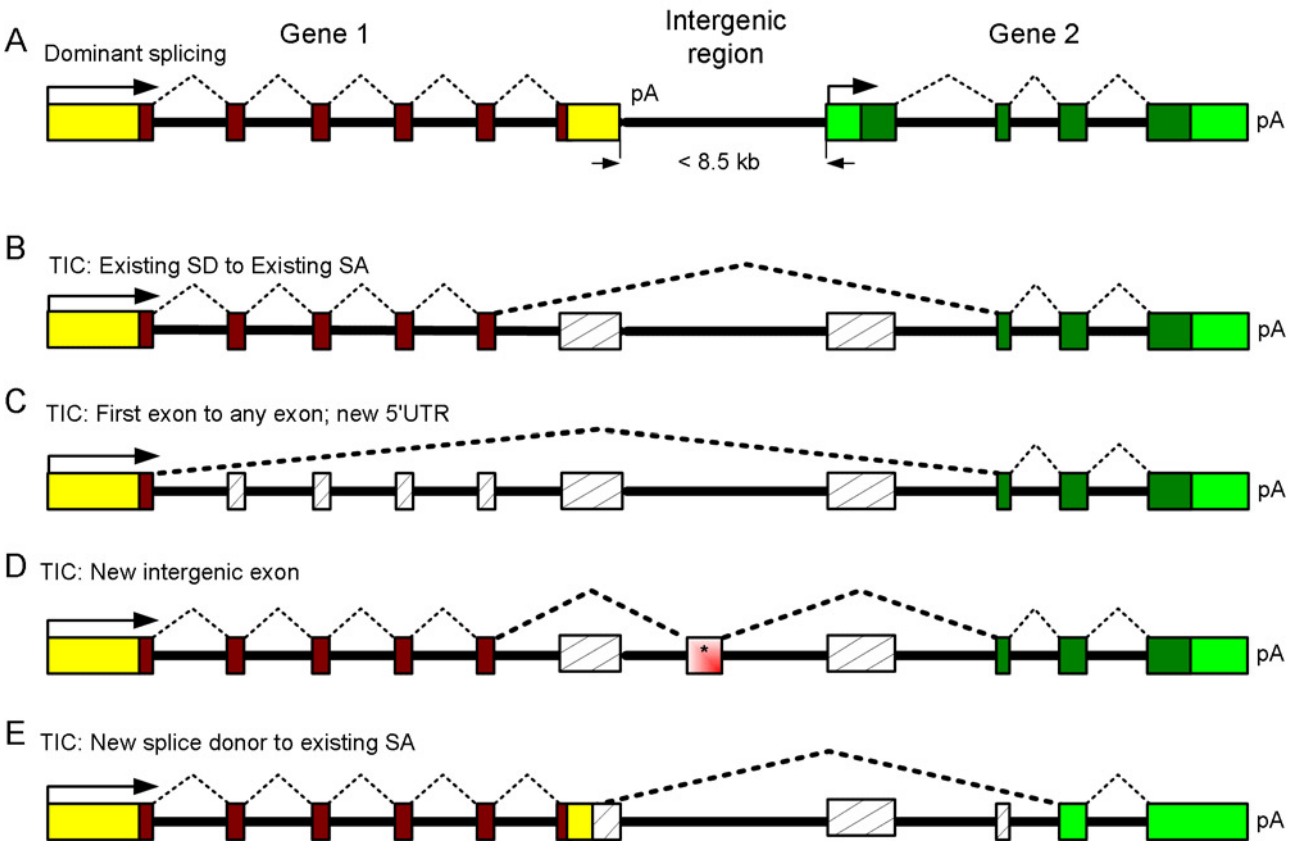

Figure 2. Co-transcription and intergenic splicing between two genes located in tandem, also known as transcription-induced chimera (TIC) [45, 46]. (*a*) The majority of TIC splicing occurs in the context of tandem genes separated by less than 8.5 kb. (*b*) The predominant TIC splicing occurs between an existing splice donor (SD) of the first gene and an existing splice acceptor (SA) in the second gene, resulting in omission of intervening exons and intergenic sequence. This generally results in fusion of the two coding sequences, out-of-frame in ~ 56% of TIC events. (*c*) Alternatively, a TIC event may combine a 5'UTR region of the first gene with the coding sequence of the second gene. (*d*) Occasionally, a novel intergenic exon may be included, or (*e*) a new splice site may be used within a framework of the existing exons. Exons are shown as rectangles, with yellow and brown shading indicating untranslated and coding regions of the first gene, and light green and dark green indicating untranslated and coding regions of the second gene, respectively. Skipped and alternatively included exons are indicated by stripes and red shading (*), respectively. Bent arrow, transcription start site; pA, polyadenylation signal; dotted lines between exons, splicing.

bidopsis thaliana [70], and Tetraodon [71]; up to two genome duplications were also proposed to occur prior to the origin of vertebrate lineage [69]. The majority of duplication events, however, involve smaller duplications either locally, creating tandem segmental duplications, or to more distant regions, typically adding to the complex pericentromeric or subtelomeric repositories of interspersed segmental duplications [1, 72, 73].

Tandem clusters of genes or duplications of exons are thought to be formed by unequal crossing-over events, or misaligned homologous recombinational repair [74, 75] (see Fig. 3); some of the genes formed by this duplicative process include the well-known examples of homeobox [76] and globin [77] genes. Similar recombination events, provoked by pairing between interspersed Alu repeat sequences, are thought to cause the formation of over 30% of interspersed segmental duplications [78, 79]. The nonuniform distribution of interspersed segmental duplications, with the highest densities in the recombinogenic pericentromeric and subtelomeric regions [1], suggest that they may be formed by illegitimate recombination and non-homologous end-joining repair processes. Reports of integration of exogenous DNA sequences at sites with little or no homology both in cultured cells and *in vivo*, often triggered by DNA damage, further support this explanation [80–82]. DNA duplicated by this mechanisms includes both exonic, intronic and intergenic sequences, and is distinct from the individual copies of processed mRNAs that are reverse transcribed and inserted into the genome by LINE-1 retrotransposons (discussed in the next section).

Segmental duplications constitute at least 5% of human genome sequence [1, 73], and are thought to arise in eukaryotes as frequently as 0.01 per gene per million years, with rates in individual species ranging

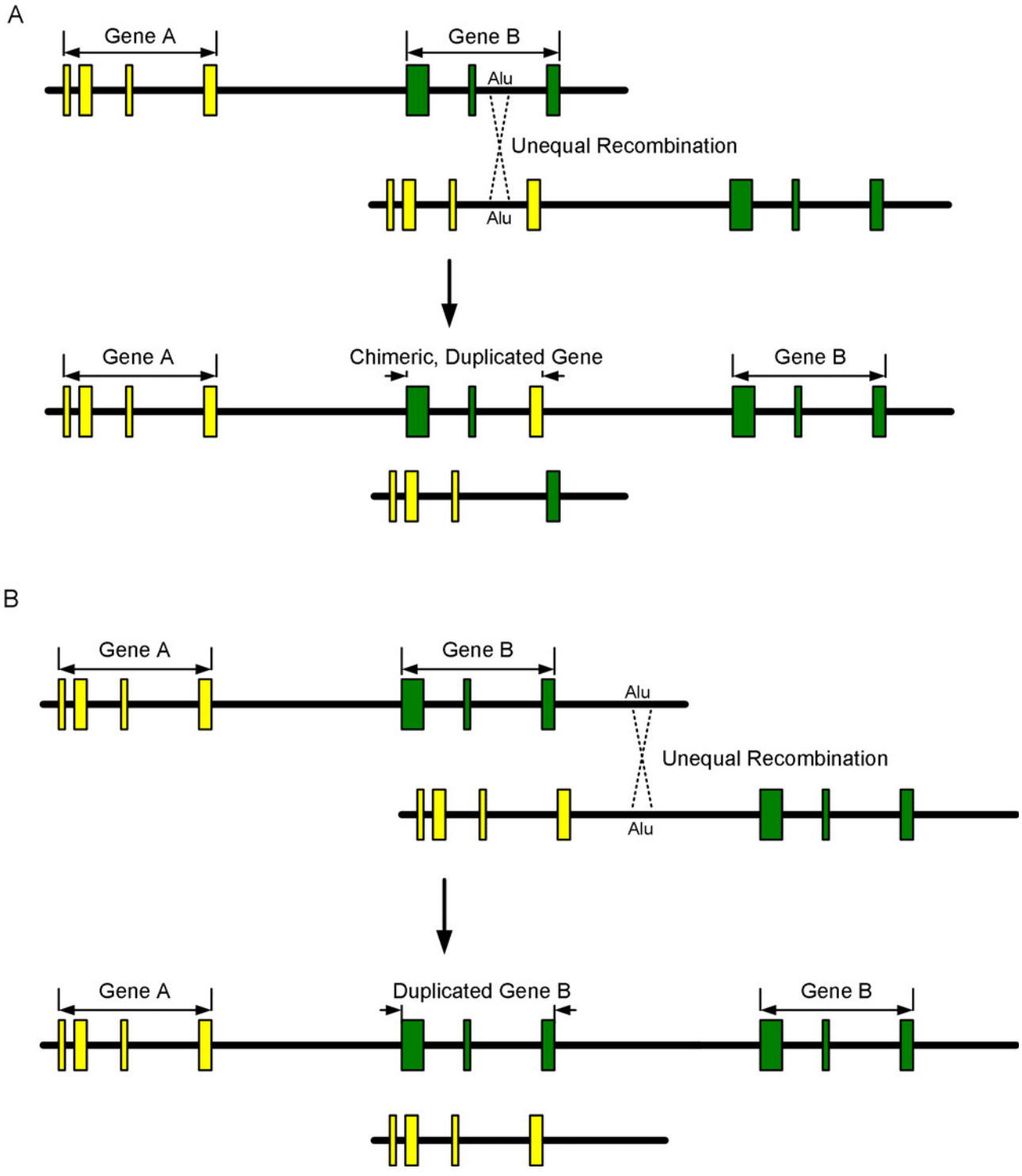

Figure 3. Unequal recombination, often mediated by homology between interspersed repeats, may lead to tandem or interspersed segmental duplications. (*a*) Unequal recombination between intronic regions of different genes may create chimeric genes and concomitant deletion products. (*b*) Recombination between intergenic regions can duplicate or delete complete gene structures. Exons are shown as rectangles, with yellow and green shading indicating gene A and B, respectively. Unequal recombination event, facilitated by Alu interspersed repeat (Alu) is shown by crossed dotted lines. Over 30% of unequal homologous recombination events are thought to occur by crossovers between Alu elements; however, nonhomologous recombination between both related and unrelated genes not facilitated by interspersed repeat sequences has also been reported [132]. The deletion product of the unequal recombination is more likely to cause haploinsufficiency, disease, or loss of viability, and be lost from the population.

100-fold from 0.002 to 0.02 [83]. Typical segmental duplications copy regions are of a few to 75 kb in length and often contain only partial gene structures insufficient for gene expression, producing dead gene copies, referred to as non-processed pseudogenes to distinguish them from processed pseudogenes created by L1 retrotransposition [72, 84].

Segmental duplications, typically clustered at the pericentromeric and subtelomeric regions, undergo subsequent rearrangement to form highly complex patterns of duplications within duplications; this is thought to be due to the active recombinogenic processes as well as lower adverse effects of integration and rearrangement in these genomic regions [79, 84, 85]. Juxtapositions of duplicons of different genes within such segments give rise to chimeric transcripts combining distinct functional domains, a few of which have maintained an open reading frame [84–86]. Thus, while segmental duplications typically lead to the formation of non-processed pseudogenes, they occasionally form gene copies that are expressed, functional, and innovative. Over the course of evolution, the sheer volume of duplicative events caused great expansion of gene families and significantly enhanced genome complexity. Particularly abundant are expansions of gene families involved in immunity and defense, membrane surface interaction, growth and development, and drug detoxification [73].

Retrotransposition

In addition to segmental duplications of whole gene structures, mammalian genomes contain over 4000 intronless copies of cellular genes, known as processed pseudogenes [87–91]. Processed pseudogenes are produced by the action of LINE-1 retrotransposon (L1, for review see [92]), which is known to occasionally reverse transcribe and integrate cellular transcripts at roughly random genomic sites [93, 94]. Early mammalian and primate lineages experienced several bursts in L1 activity creating large populations of processed pseudogenes [89].

Because L1 retrotransposes typical pol II transcripts that were spliced and polyadenylated, the resulting processed pseudogenes, unlike gene copies in segmental duplications, are intronless and lack the original promoter and regulatory sequences of template mRNA (see Fig. 4). For this reason, processed pseudogenes have been classically thought to be transcriptionally dead on arrival; the only exceptions were those arising from alternative transcripts that aberrantly included the gene's original promoter, or pseudogenes that happened to integrate within an existing transcription unit [95–98]. However, more recent genome-wide surveys of EST databases as well as transcription analyses of individual pseudogenes have revealed that, in fact, up to a third of processed pseudogenes are transcribed, most of them specifically in the testes [90, 95, 97, 99–104].

The testicular specificity of pseudogene transcription is thought to be caused by the transcriptionally permissive environment in the testes, where the components of the polymerase II (pol II) holoenzyme complex are known to increase 30–1000-fold during the haploid stages of spermatogenesis [105, 106]. While in other tissues transcription initiation requires transcription factor recruitment that is inefficient in pseudogenes, the greatly increased concentrations of holoenzyme in the testes allow for efficient transcription from otherwise suboptimal promoter sequences [107, 108]. It has been proposed that testicular expression allows for the initial functional adaptation of the pseudogene in the testes and is likely to contribute to male reproductive evolution and speciation; occasional acquisition of more diverse regulatory elements by mutation may subsequently allow broader expression [99].

Dozens of functional pseudogenes (known as retrogenes) have been described, and the total number of retrogenes in the human genome is estimated at ~120 [99]. Among them, there is a striking predominance of autosomal retrogenes, which are copies of X-linked parental genes [99, 109, 110]. This phenomenon has been attributed to the functional substitution of the autosomal retrogene for the X-linked parental gene; X-linked genes are silenced during spermatogenesis by male X chromosome inactivation [95, 111–113]. A remarkable example of retrogene compensation for the silencing of the X-linked homologous gene was described by Bradley and colleagues [98]. They mapped the mutation causing juvenile spermatogonial depletion (jsd) in mice to autosomal Utp14b retrogene, which arose by retrotransposition of the X-linked conserved gene Utp14a, a mammalian ortholog of yeast Utp14 protein required for pre-rRNA processing and ribosome assembly. Utp14b is expressed primarily in the testes and is thought to substitute for the silenced Utp14a in spermatogenesis. Using sequence-based phylogenetic analyses, Bradley et al. [98] found that, while Utp14b arose in mice after the rodent-primate split, an independent retrotransposition event of Utp14a created a homologous UTP14C retrogene in humans, likely driven by the selective pressure to support spermatogenesis during meiotic X inactivation.

Unlike segmental duplications, known to cluster in limited recombinogenic regions [1, 72], retrotransposition is roughly random, frequently inserting retrogenes in close vicinity to or within other genes, thus

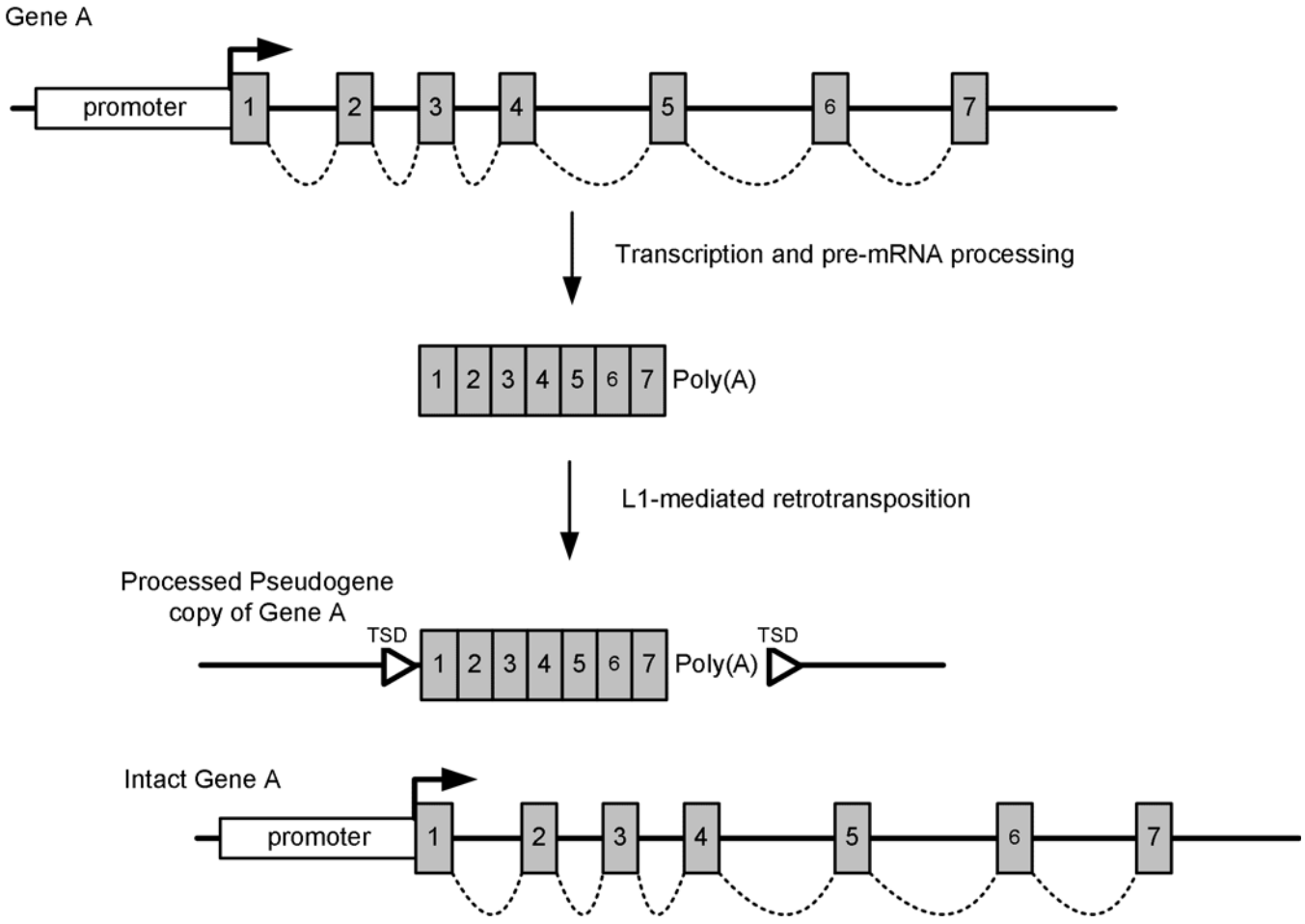

Figure 4. L1 retrotransposon mobilizes processed mRNAs of cellular genes to create processed pseudogenes. Typical pre-mRNA transcribed by RNA polymerase II begins downstream of its promoter sequence. Following transcription, pre-mRNAs are spliced, m7G-capped and polyadenylated. Occasionally, a processed transcript is reverse transcribed and integrated into the genome by L1 retrotransposon (reviewed in [92]), forming an intronless copy of the original gene in a new location. L1-mobilized intronless gene copies are known as processed pseudogenes; they typically end in a poly(A) sequence, and are flanked by short direct repeats (also known as target site duplications or TSDs). Gray rectangles, exons; white rectangle, promoter sequence; bent arrow, transcription start site; dotted lines, splicing; open triangles, TSDs.

creating a fertile environment for exon shuffling. An important example of genome innovation by this mechanism includes TRIMCyp gene, formed by L1 retrotransposition of cyclophilin A (CypA) transcript into intron 7 of TRIM5 ubiquitin ligase in owl monkey [114]. Sayah and colleagues [114] discovered that the intronic CypA insertion is spliced to TRIM5 exon 5, causing an in-frame fusion of TRIM5 (exons 1–7) with the complete CypA cDNA; the resulting chimeric protein confers HIV-1 resistance in owl monkey. Frequent co-transcription and retrotransposition of 3' flanking DNA by L1 retrotransposons contributes to exon-shuffling by a similar mechanism, whereby L1 may transfer its downstream flanking region (genic or nongenic) next to an existing gene, where it may then develop into a new exon [115, 116].

Notably, retrotransposition plays a key and unique role in increasing gene architectures by enabling the transformation of low-abundance innovative transcripts into retrogenes that can be inherited, expressed, and can evolve independently from the original gene locus. Indeed, genome-wide comparisons of processed pseudogenes with EST and cDNA databases demonstrate that 19% of pseudogenes represent unique ancestral splice variants, representing alternatively spliced as well as TIC mRNAs that are rare or not found in current humans [117]. It is conceivable that even occasional transcripts formed by normally rare trans-splicing [118–120] or RNA-RNA recombination [121, 122] between distinct genes may have been retrotransposed and contributed to domain shuffling and gene expansion over the course of eukaryotic evolution.

In addition to faithful retro-duplication of canonical and innovative cellular transcripts, L1 has generated a number of chimeric pseudogenes, combining portions

of different mRNAs or mRNAs and transposable elements, by switching of RNA templates during reverse transcription [123–125]. The majority of these chimeric events involve spliceosomal snRNAs U6 and U5 and rRNA-processing snoRNA U3 in combination with L1 or Alu elements, likely reflecting the close spatial relationships of these RNAs and the L1 retrotransposition complex within the cell [124, 126]. Similar template switching events could have contributed to the co-opting of transposable element sequences for cellular gene functions [127].

Gene creation and loss are frequent, ongoing, and are key contributors to speciation

In sum, the formation of novel genes is frequent, ubiquitous, and ongoing, with identified cases being but a small fraction of total events: those recent enough to be recognizable, yet old enough to be fixed or present at a high enough frequency in the population to be found in sequenced genomes and EST databases. Faithful gene copies, created by segmental duplications or retrotransposition, start out with redundant function to their parental gene, and are subject to relaxed functional constraint and neutral evolution [128]; an intriguing exception are retrogenes with essential functions in spermatogenesis [98]. In contrast, chimeric genes, created by either tandem duplications within an existing gene, juxtaposition of segmental duplications, or by retrotransposition, begin with a unique function and are thought to undergo a burst of positive Darwinian selection, followed by quiescence and increasing functional constraint [129]. Relaxed functional constraint and rapid evolution, in turn, lead to the rapid demise of the vast majority of gene duplicates through mutational inactivation, with an estimated half life of ~4 Myr [83]. Thus, frequent gene duplication followed by differential loss and, more rarely, acquisition of differential function is likely a significant force in the evolution of genome complexity, reproductive isolation and the divergence of species [7, 74, 83].

Acknowledgements. We would like to thank members of the Kazazian lab for scientific discussions, and two expert anonymous referees for their insightful and helpful suggestions that improved this manuscript. This work was supported by a grant from National Institutes of Health to H.H.K.

1 Lander E. S., Linton L. M., Birren B., Nusbaum C., Zody M. C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., Funke R., Gage D. et al. (2001) Initial sequencing and analysis of the human genome. Nature 409, 860 – 921.
2 Gibbs R. A., Weinstock G. M., Metzker M. L., Muzny D. M., Sodergren E. J., Scherer S., Scott G., Steffen D., Worley K. C., Burch P. E., Okwuonu G., Hines S. et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428, 493 – 521.
3 Waterston R. H., Lindblad-Toh K., Birney E., Rogers J., Abril J. F., Agarwal P., Agarwala R., Ainscough R., Alexandersson M., An P., Antonarakis S. E., Attwood J. et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520 – 562.
4 Lindblad-Toh K., Wade C. M., Mikkelsen T. S., Karlsson E. K., Jaffe D. B., Kamal M., Clamp M., Chang J. L., Kulbokas E. J., 3rd, Zody M. C., Mauceli E., Xie X. et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438, 803 – 819.
5 The Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437, 69 – 87.
6 Venter J. C., Adams M. D., Myers E. W., Li P. W., Mural R. J., Sutton G. G., Smith H. O., Yandell M., Evans C. A., Holt R. A., Gocayne J. D., Amanatides P. et al. (2001) The sequence of the human genome. Science 291, 1304 – 1351.
7 Hillier L. W., Miller W., Birney E., Warren W., Hardison R. C., Ponting C. P., Bork P., Burt D. W., Groenen M. A., Delany M. E., Dodgson J. B., Chinwalla A. T. et al. (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432, 695 – 716.
8 Aparicio S., Chapman J., Stupka E., Putnam N., Chia J. M., Dehal P., Christoffels A., Rash S., Hoon S., Smit A., Gelpke M. D., Roach J. et al. (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297, 1301 – 1310.
9 Liolios K., Tavernarakis N., Hugenholtz P. and Kyrpides N. C. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. Nucleic Acids Res. 34, D332 – 334.
10 Birney E., Andrews D., Caccamo M., Chen Y., Clarke L., Coates G., Cox T., Cunningham F., Curwen V., Cutts T., Down T., Durbin R. et al. (2006) Ensembl 2006. Nucleic Acids Res. 34, D556 – 561.
11 NCBI, http://www.ncbi.nlm.nih.gov/
12 Furey T. S. (2006) Comparison of human (and other) genome browsers. Hum. Genomics 2, 266 – 270.
13 Karolchik D., Baertsch R., Diekhans M., Furey T. S., Hinrichs A., Lu Y. T., Roskin K. M., Schwartz M., Sugnet C. W., Thomas D. J., Weber R. J., Haussler D. and Kent W. J. (2003) The UCSC Genome Browser Database. Nucleic Acids Res. 31, 51 – 54.
14 Lander E. S. (1996) The new genomics: global views of biology. Science 274, 536 – 539.
15 Thomas J. W., Touchman J. W., Blakesley R. W., Bouffard G. G., Beckstrom-Sternberg S. M., Margulies E. H., Blanchette M., Siepel A. C., Thomas P. J., McDowell J. C., Maskeri B., Hansen N. F. et al. (2003) Comparative analyses of multispecies sequences from targeted genomic regions. Nature 424, 788 – 793.
16 Margulies E. H., Maduro V. V., Thomas P. J., Tomkins J. P., Amemiya C. T., Luo M. and Green E. D. (2005) Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes. Proc. Natl. Acad. Sci. USA 102, 3354 – 3359.
17 Darwin C. (1859) By Means Of Natural Selection, Or The Preservation Of Favoured Races In The Struggle For Life. John Murray, London
18 Taylor J. S. and Raes J. (2004) Duplication and divergence: the evolution of new genes and old ideas. Annu. Rev. Genet. 38, 615 – 643.
19 Dorit R. L., Schoenbach L. and Gilbert W. (1990) How big is the universe of exons? Science 250, 1377 – 1382.
20 Li W. H., Gu Z., Wang H. and Nekrutenko A. (2001) Evolutionary analyses of the human genome. Nature 409, 847 – 849.

21 Long M., Rosenberg C. and Gilbert W. (1995) Intron phase correlations and the evolution of the intron/exon structure of genes. Proc. Natl. Acad. Sci. USA 92, 12495 – 12499.
22 de Souza S. J., Long M., Schoenbach L., Roy S. W. and Gilbert W. (1996) Intron positions correlate with module boundaries in ancient proteins. Proc. Natl. Acad. Sci. USA 93, 14632 – 14636.
23 Kaessmann H., Zollner S., Nekrutenko A. and Li W. H. (2002) Signatures of domain shuffling in the human genome. Genome Res. 12, 1642 – 1650.
24 Vibranovski M. D., Sakabe N. J., de Oliveira R. S. and de Souza S. J. (2005) Signs of ancient and modern exon-shuffling are correlated to the distribution of ancient and modern domains along proteins. J. Mol. Evol. 61, 341 – 350.
25 Patthy L. (1996) Exon shuffling and other ways of module exchange. Matrix Biol. 15, 301 – 310.
26 Patthy L. (2003) Modular assembly of genes and the evolution of new functions. Genetica 118, 217 – 231.
27 Claverie J. M. (2001) Gene number. What if there are only 30,000 human genes? Science 291, 1255 – 1257.
28 Long M., Betran E., Thornton K. and Wang W. (2003) The origin of new genes: glimpses from the young and old. Nat. Rev. Genet. 4, 865 – 875.
29 Salzberg S. L., White O., Peterson J. and Eisen J. A. (2001) Microbial genes in the human genome: Lateral transfer or gene loss? Science 292, 1903 – 1906.
30 Stanhope M. J., Lupas A., Italia M. J., Koretke K. K., Volker C. and Brown J. R. (2001) Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. Nature 411, 940 – 944.
31 DeFilippis V., Villarreal; L. P., Salzberg S. L. and Eisen J. A. (2001) Lateral gene transfer or viral colonization? Science 293, 1048a
32 Roelofs J. and Van Haastert P. J. (2001) Genes lost during evolution. Nature 411, 1013 – 1014.
33 Andersson J. O. (2005) Lateral gene transfer in eukaryotes. Cell. Mol. Life Sci. 62, 1182 – 1197.
34 Modrek B., Resch A., Grasso C. and Lee C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res. 29, 2850 – 2859.
35 Johnson J. M., Castle J., Garrett-Engele P., Kan Z., Loerch P. M., Armour C. D., Santos R., Schadt E. E., Stoughton R. and Shoemaker D. D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302, 2141 – 2144.
36 Stolc V., Gauhar Z., Mason C., Halasz G., van Batenburg M. F., Rifkin S. A., Hua S., Herreman T., Tongprasit W., Barbano P. E., Bussemaker H. J. and White K. P. (2004) A gene expression map for the euchromatic genome of *Drosophila melanogaster*. Science 306, 655 – 660.
37 Blencowe B. J. (2006) Alternative splicing: new insights from global analyses. Cell 126, 37 – 47.
38 Resch A., Xing Y., Modrek B., Gorlick M., Riley R. and Lee C. (2004) Assessing the impact of alternative splicing on domain interactions in the human proteome. J. Proteome Res. 3, 76 – 83.
39 Yeo G., Holste D., Kreiman G. and Burge C. B. (2004) Variation in alternative splicing across human tissues. Genome Biol. 5, R74.
40 Crawford J. B. and Patton J. G. (2006) Activation of alpha - tropomyosin exon 2 is regulated by the SR protein 9G8 and heterogeneous nuclear ribonucleoproteins H and F. Mol. Cell. Biol. 26, 8791 – 8802.
41 Chaudhuri T., Mukherjea M., Sachdev S., Randall J. D. and Sarkar S. (2005) Role of the fetal and alpha/beta exons in the function of fast skeletal troponin T isoforms: correlation with altered Ca^{2+} regulation associated with development. J. Mol. Biol. 352, 58 – 71.
42 Kogerman P., Krause D., Rahnama F., Kogerman L., Unden A. B., Zaphiropoulos P. G. and Toftgard R. (2002) Alternative first exons of PTCH1 are differentially regulated *in vivo* and may confer different functions to the PTCH1 protein. Oncogene 21, 6007 – 6016.
43 Zaphiropoulos P. G. (1999) RNA molecules containing exons originating from different members of the cytochrome P450 2C gene subfamily (CYP2C) in human epidermis and liver. Nucleic Acids Res. 27, 2585 – 2590.
44 Magrangeas F., Pitiot G., Dubois S., Bragado-Nilsson E., Cherel M., Jobert S., Lebeau B., Boisteau O., Lethe B., Mallet J., Jacques Y. and Minvielle S. (1998) Cotranscription and intergenic splicing of human galactose-1-phosphate uridylyltransferase and interleukin-11 receptor alpha-chain genes generate a fusion mRNA in normal cells. Implication for the production of multidomain proteins during evolution. J. Biol. Chem. 273, 16005 – 16010.
45 Akiva P., Toporik A., Edelheit S., Peretz Y., Diber A., Shemesh R., Novik A. and Sorek R. (2006) Transcription-mediated gene fusion in the human genome. Genome Res. 16, 30 – 36.
46 Parra G., Reymond A., Dabbouseh N., Dermitzakis E. T., Castelo R., Thomson T. M., Antonarakis S. E. and Guigo R. (2006) Tandem chimerism as a means to increase protein complexity in the human genome. Genome Res. 16, 37 – 44.
47 Fears S., Mathieu C., Zeleznik-Le N., Huang S., Rowley J. D. and Nucifora G. (1996) Intergenic splicing of MDS1 and EVI1 occurs in normal tissues as well as in myeloid leukemia and produces a new member of the PR domain family. Proc. Natl. Acad. Sci. USA 93, 1642 – 1647.
48 Pardigol A., Forssmann U., Zucht H. D., Loetscher P., Schulz-Knappe P., Baggiolini M., Forssmann W. G. and Magert H. J. (1998) HCC-2, a human chemokine: gene structure, expression pattern, and biological activity. Proc. Natl. Acad. Sci. USA 95, 6308 – 6313.
49 Upadhyaya A. B., Lee S. H. and DeJong J. (1999) Identification of a general transcription factor TFIIAalpha/beta homolog selectively expressed in testis. J. Biol. Chem. 274, 18040 – 18048.
50 Millar J. K., Christie S., Semple C. A. and Porteous D. J. (2000) Chromosomal location and genomic structure of the human translin-associated factor X gene (TRAX; TSNAX) revealed by intergenic splicing to DISC1, a gene disrupted by a translocation segregating with schizophrenia. Genomics 67, 69 – 77.
51 Thomson T. M., Lozano J. J., Loukili N., Carrio R., Serras F., Cormand B., Valeri M., Diaz V. M., Abril J., Burset M., Merino J., Macaya A., Corominas M. and Guigo R. (2000) Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene. Genome Res. 10, 1743 – 1756.
52 Communi D., Suarez-Huerta N., Dussossoy D., Savi P. and Boeynaems J. M. (2001) Cotranscription and intergenic splicing of human P2Y11 and SSF1 genes. J. Biol. Chem. 276, 16561 – 16566.
53 Cox P. R., Siddique T. and Zoghbi H. Y. (2001) Genomic organization of Tropomodulins 2 and 4 and unusual intergenic and intraexonic splicing of YL-1 and Tropomodulin 4. BMC Genomics 2, 7.
54 Kolfschoten G. M., Pradet-Balade B., Hahne M. and Medema J. P. (2003) TWE-PRIL; a fusion protein of TWEAK and APRIL. Biochem. Pharmacol. 66, 1427 – 1432.
55 Kato M., Khan S., Gonzalez N., O'Neill B. P., McDonald K. J., Cooper B. J., Angel N. Z. and Hart D. N. (2003) Hodgkin's lymphoma cell lines express a fusion protein encoded by intergenically spliced mRNA for the multilectin receptor DEC-205 (CD205) and a novel C-type lectin receptor DCL-1. J. Biol. Chem. 278, 34035 – 34041.
56 Poulin F., Brueschke A. and Sonenberg N. (2003) Gene fusion and overlapping reading frames in the mammalian genes for 4E-BP3 and MASK. J. Biol. Chem. 278, 52290 – 52297.
57 Maeda K., Horikoshi T., Nakashima E., Miyamoto Y., Mabuchi A. and Ikegawa S. (2005) MATN and LAPTM are parts of larger transcription units produced by intergenic

splicing: Intergenic splicing may be a common phenomenon. DNA Res. 12, 365 – 372.
58 Roux M., Leveziel H. and Amarger V. (2006) Cotranscription and intergenic splicing of the PPARG and TSEN2 genes in cattle. BMC Genomics 7, 71.
59 Nekrutenko A. (2004) Identification of novel exons from rat-mouse comparisons. J. Mol. Evol. 59, 703 – 708.
60 Wang W., Zheng H., Yang S., Yu H., Li J., Jiang H., Su J., Yang L., Zhang J., McDermott J., Samudrala R., Wang J., Yang H., Yu J., Kristiansen K. and Wong G. K. (2005) Origin and evolution of new exons in rodents. Genome Res. 15, 1258 – 1264.
61 Makalowski W., Mitchell G. A. and Labuda D. (1994) Alu sequences in the coding regions of mRNA: a source of protein variability. Trends Genet. 10, 188 – 193.
62 Nekrutenko A. and Li W. H. (2001) Transposable elements are found in a large number of human protein-coding genes. Trends Genet. 17, 619 – 621.
63 Zhang X. H. and Chasin L. A. (2006) Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. Proc. Natl. Acad. Sci. USA 103, 13427 – 13432.
64 Gotea V. and Makalowski W. (2006) Do transposable elements really contribute to proteomes? Trends Genet. 22, 260 – 267.
65 Pavlicek A., Clay O. and Bernardi G. (2002) Transposable elements encoding functional proteins: pitfalls in unprocessed genomic data? FEBS Lett. 523, 252 – 253.
66 Belancio V. P., Hedges D. J. and Deininger P. (2006) LINE-1 RNA splicing and influences on mammalian gene expression. Nucleic Acids Res. 34, 1512 – 1521.
67 Lewis B. P., Green R. E. and Brenner S. E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc. Natl. Acad. Sci. USA 100, 189 – 192.
68 Nagy E. and Maquat L. E. (1998) A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends Biochem. Sci. 23, 198 – 199.
69 McLysaght A., Hokamp K. and Wolfe K. H. (2002) Extensive genomic duplication during early chordate evolution. Nat. Genet. 31, 200 – 204.
70 Blanc G. and Wolfe K. H. (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16, 1679 – 1691.
71 Jaillon O., Aury J. M., Brunet F., Petit J. L., Stange-Thomann N., Mauceli E., Bouneau L, Fischer C., Ozouf-Costaz C., Bernot A., Nicaud S., Jaffe D. et al. (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431, 946 – 957.
72 Samonte R. V. and Eichler E. E. (2002) Segmental duplications and the evolution of the primate genome. Nat. Rev. Genet. 3, 65 – 72.
73 Bailey J. A., Gu Z., Clark R. A., Reinert K., Samonte R. V., Schwartz S., Adams M. D., Myers E. W., Li P. W. and Eichler E. E. (2002) Recent segmental duplications in the human genome. Science 297, 1003 – 1007.
74 Zhang J. (2003) Evolution by gene duplication: an update. Trends Ecol. Evol. 18, 292 – 298.
75 Thompson L. H. and Schild D. (2001) Homologous recombinational repair of DNA ensures mammalian chromosome stability. Mutat Res. 477, 131 – 153.
76 Holland P. W. and Takahashi T. (2005) The evolution of homeobox genes: Implications for the study of brain development. Brain Res Bull. 66, 484 – 490.
77 Shen S. H., Slightom J. L. and Smithies O. (1981) A history of the human fetal globin gene duplication. Cell 26, 191 – 203.
78 Bailey J. A., Liu G. and Eichler E. E. (2003) An Alu transposition model for the origin and expansion of human segmental duplications. Am. J. Hum. Genet. 73, 823 – 834.
79 Babcock M., Pavlicek A., Spiteri E., Kashork C. D., Ioshikhes I., Shaffer L. G., Jurka J. and Morrow B. E. (2003) Shuffling of genes within low-copy repeats on 22q11 (LCR22) by Alu-mediated recombination events during evolution. Genome Res. 13, 2519 – 2532.
80 van Rijk A. and Bloemendal H. (2003) Molecular mechanisms of exon shuffling: illegitimate recombination. Genetica 118, 245 – 249.
81 van Rijk A. A., de Jong, W. W. and Bloemendal, H. (1999) Exon shuffling mimicked in cell culture. Proc. Natl. Acad. Sci. USA 96, 8074 – 8079.
82 Roth D. B., Porter T. N. and Wilson J. H. (1985) Mechanisms of nonhomologous recombination in mammalian cells. Mol. Cell. Biol. 5, 2599 – 2607.
83 Lynch M. and Conery J. S. (2000) The evolutionary fate and consequences of duplicate genes. Science 290, 1151 – 1155.
84 Eichler E. E. (2001) Recent duplication, domain accretion and the dynamic mutation of the human genome. Trends Genet. 17, 661 – 669.
85 Stankiewicz P., Shaw C. J., Withers M., Inoue K. and Lupski J. R. (2004) Serial segmental duplications during primate evolution result in complex human genome architecture. Genome Res. 14, 2209 – 2220.
86 Courseaux A., Richard F., Grosgeorge J., Ortola C., Viale A., Turc-Carel C., Dutrillaux B., Gaudray P. and Nahon J.-L. (2003) Segmental duplications in euchromatic regions of human chromosome 5: A source of evolutionary instability and transcriptional innovation. Genome Res. 13, 369 – 381.
87 Zhang Z., Harrison P. M., Liu Y. and Gerstein M. (2003) Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res. 13, 2541 – 2558.
88 Zhang Z., Carriero N. and Gerstein M. (2004) Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet. 20, 62 – 67.
89 Ohshima K., Hattori M., Yada T., Gojobori T., Sakaki Y. and Okada N. (2003) Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. Genome Biol. 4, R74.
90 Marques A. C., Dupanloup I., Vinckenbosch N., Reymond A. and Kaessmann H. (2005) Emergence of young human genes after a burst of retroposition in primates. PLoS Biol. 3, e357.
91 Khelifi A., Duret L. and Mouchiroud D. (2005) HOPPSIGEN: a database of human and mouse processed pseudogenes. Nucleic Acids Res. 33, D59 – 66.
92 Ostertag E. M. and Kazazian H. H. Jr. (2001) Biology of mammalian L1 retrotransposons. Annu. Rev. Genet. 35, 501 – 538.
93 Wei W., Gilbert N., Ooi S. L., Lawler J. F., Ostertag E. M., Kazazian H. H., Boeke J. D. and Moran J. V. (2001) Human L1 retrotransposition: cis preference versus trans complementation. Mol. Cell. Biol. 21, 1429 – 1439.
94 Esnault C., Maestre J. and Heidmann T. (2000) Human LINE retrotransposons generate processed pseudogenes. Nat. Genet. 24, 363 – 367.
95 McCarrey J. R. and Thomas K. (1987) Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. Nature 326, 501 – 505.
96 Mighell A. J., Smith N. R., Robinson P. A. and Markham A. F. (2000) Vertebrate pseudogenes. FEBS Lett. 468, 109 – 114.
97 Makeyev A. V., Chkheidze A. N. and Liebhaber S. A. (1999) A set of highly conserved RNA-binding proteins, alphaCP-1 and alphaCP-2, implicated in mRNA stabilization, are coexpressed from an intronless gene and its intron-containing paralog. J. Biol. Chem. 274, 24849 – 24857.
98 Bradley J., Baltus A., Skaletsky H., Royce-Tolland M., Dewar K. and Page D. C. (2004) An X-to-autosome retrogene is required for spermatogenesis in mice. Nat. Genet. 36, 872 – 876.
99 Vinckenbosch N., Dupanloup I. and Kaessmann H. (2006) Evolutionary fate of retroposed gene copies in the human genome. Proc. Natl. Acad. Sci. USA 103, 3220 – 3225.

100 Kleene K. C., Mulligan E., Steiger D., Donohue K. and Mastrangelo M. A. (1998) The mouse gene encoding the testis-specific isoform of Poly(A) binding protein (Pabp2) is an expressed retroposon: intimations that gene expression in spermatogenic cells facilitates the creation of new genes. J. Mol. Evol. 47, 275 – 281.
101 Betran E. and Long M. (2003) Dntf-2r, a young Drosophila retroposed gene with specific male expression under positive Darwinian selection. Genetics 164, 977 – 988.
102 Harrison P. M., Zheng D., Zhang Z., Carriero N. and Gerstein M. (2005) Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. Nucleic Acids Res. 33, 2374 – 2383.
103 Wang P. J. and Page D. C. (2002) Functional substitution for TAF(II)250 by a retroposed homolog that is expressed in human spermatogenesis. Hum. Mol. Genet. 11, 2341 – 2346.
104 Burki F. and Kaessmann H. (2004) Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. Nat. Genet. 36, 1061 – 1063.
105 Schmidt E. E. and Schibler U. (1995) High accumulation of components of the RNA polymerase II transcription machinery in rodent spermatids. Development 121, 2373 – 2383.
106 Schmidt E. E. and Schibler U. (1997) Developmental testis-specific regulation of mRNA levels and mRNA translational efficiencies for TATA-binding protein mRNA isoforms. Dev. Biol. 184, 138 – 149.
107 Schmidt E. E. (1996) Transcriptional promiscuity in testes. Curr. Biol. 6, 768 – 769.
108 Ossipow V., Tassan J. P., Nigg E. A. and Schibler U. (1995) A mammalian RNA polymerase II holoenzyme containing all components required for promoter-specific transcription initiation. Cell 83, 137 – 146.
109 Emerson J. J., Kaessmann H., Betran E. and Long M. (2004) Extensive gene traffic on the mammalian X chromosome. Science 303, 537 – 540.
110 Betran E., Thornton K. and Long M. (2002) Retroposed new genes out of the X in Drosophila. Genome Res. 12, 1854 – 1859.
111 Handel M. A., Park C. and Kot M. (1994) Genetic control of sex-chromosome inactivation during male meiosis. Cytogenet. Cell Genet. 66, 83 – 88.
112 Richler C., Soreq H. and Wahrman J. (1992) X inactivation in mammalian testis is correlated with inactive X-specific transcription. Nat. Genet. 2, 192 – 195.
113 Turner J. M., Mahadevaiah S. K., Elliott D. J., Garchon H. J., Pehrson J. R., Jaenisch R. and Burgoyne P. S. (2002) Meiotic sex chromosome inactivation in male mice with targeted disruptions of Xist. J. Cell Sci. 115, 4097 – 4105.
114 Sayah D. M., Sokolskaja E., Berthoux L. and Luban J. (2004) Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. Nature 430, 569 – 573.
115 Goodier J. L., Ostertag E. M. and Kazazian H. H. Jr. (2000) Transduction of 3'-flanking sequences is common in L1 retrotransposition. Hum. Mol. Genet. 9, 653 – 657.
116 Moran J. V., DeBerardinis R. J. and Kazazian H. H. J. (1999) Exon shuffling by L1 retrotransposition. Science 283, 1530 – 1534.
117 Shemesh R., Novik A., Edelheit S. and Sorek R. (2006) Genomic fossils as a snapshot of the human transcriptome. Proc. Natl. Acad. Sci. USA 103, 1364 – 1369.
118 Finta C. and Zaphiropoulos P. G. (2002) Intergenic mRNA molecules resulting from trans-splicing. J. Biol. Chem. 277, 5882 – 5890.
119 Caudevilla C., Serra D., Miliar A., Codony C., Asins G., Bach M. and Hegardt F. G. (1998) Natural trans-splicing in carnitine octanoyltransferase pre-mRNAs in rat liver. Proc. Natl. Acad. Sci. USA. 95, 12185 – 12190.
120 Pasman Z. and Garcia-Blanco M. A. (1996) The 5' and 3' splice sites come together via a three dimensional diffusion mechanism. Nucleic Acids Res. 24, 1638 – 1645.
121 Chetverina H. V., Demidenko A. A., Ugarov V. I. and Chetverin A. B. (1999) Spontaneous rearrangements in RNA sequences. FEBS Lett. 450, 89 – 94.
122 Gmyl A. P., Belousov E. V., Maslova S. V., Khitrina E. V., Chetverin A. B. and Agol V. I. (1999) Nonreplicative RNA recombination in poliovirus. J. Virol. 73, 8958 – 8965.
123 Buzdin A., Ustyugova S., Gogvadze E., Vinogradova T., Lebedev Y. and Sverdlov E. (2002) A new family of chimeric retrotranscripts formed by a full copy of U6 small nuclear RNA fused to the 3' terminus of l1. Genomics 80, 402 – 406.
124 Buzdin A., Gogvadze E., Kovalskaya E., Volchkov P., Ustyugova S., Illarionova A., Fushan A., Vinogradova T. and Sverdlov E. (2003) The human genome contains many types of chimeric retrogenes generated through *in vivo* RNA recombination. Nucleic Acids Res. 31, 4385 – 4390.
125 Gogvadze E. V., Buzdin A. A. and Sverdlov E. D. (2005) Multiple template switches on LINE-directed reverse transcription: the most probable formation mechanism for the double and triple chimeric retroelements in mammals. Bioorg. Khim. 31, 82 – 89.
126 Goodier J. L., Ostertag E. M., Engleka K. A., Seleme M. C. and Kazazian H. H. Jr (2004) A potential role for the nucleolus in L1 retrotransposition. Hum. Mol. Genet. 13, 1041 – 1048.
127 Britten R. J. (2004) Coding sequences of functioning human genes derived entirely from mobile element sequences. Proc. Natl. Acad. Sci. USA 101, 16825 – 16830.
128 Kimura M. (1983) The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge
129 Jones C. D. and Begun D. J. (2005) Parallel evolution of chimeric fusion genes. Proc. Natl. Acad. Sci. USA 102, 11373 – 11378.
130 Rubin G. M., Yandell M. D., Wortman J. R., Gabor Miklos G. L., Nelson C. R., Hariharan I. K., Fortini M. E., Li P. W., Apweiler R., Fleischmann W., Cherry J. M., Henikoff S. et al. (2000) Comparative genomics of the eukaryotes. Science 287, 2204 – 2215.
131 Kim J. M., Vanguri S., Boeke J. D., Gabriel A. and Voytas D. F. (1998) Transposable elements and genome organization: A comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Res. 8, 464 – 478.
132 Jones J. M., Huang J. D., Mermall V., Hamilton B. A., Mooseker M. S., Escayg A., Copeland N. G., Jenkins N. A. and Meisler M. H. (2000) The mouse neurological mutant flailer expresses a novel hybrid gene derived by exon shuffling between Gnb5 and Myo5a. Hum. Mol. Genet. 9, 821 – 828.