RNA-based gene duplication: mechanistic and evolutionary insights

Henrik Kaessmann, Nicolas Vinckenbosch* and Manyuan Long‡*

Abstract | Gene copies that stem from the mRNAs of parental source genes have long been viewed as evolutionary dead-ends with little biological relevance. Here we review a range of recent studies that have unveiled a significant number of functional retroposed gene copies in both mammalian and some non-mammalian genomes. These studies have not only revealed previously unknown mechanisms for the emergence of new genes and their functions but have also provided fascinating general insights into molecular and evolutionary processes that have shaped genomes. For example, analyses of chromosomal gene movement patterns via RNA-based gene duplication have shed fresh light on the evolutionary origin and biology of our sex chromosomes.

New gene
A gene that originated recently during evolution.

Parental gene
Source of the mRNA that gives rise to a retroposed gene copy.

Retrogene
Expressed and functional retrocopy, usually with an intact ORF consistent with that of the parental gene.

Gene fusion
The fusion of adjacent genes into a single transcription unit, which is then termed a chimeric or fusion gene.

**Center for Integrative Genomics, University of Lausanne, Genopode, CH-1015 Lausanne, Switzerland.*
‡Department of Ecology and Evolution, The University of Chicago, 1101 East 57th Street, Chicago, Illinois 60637, USA.
Correspondence to H.K.
e-mail:
Henrik.Kaessmann@unil.ch
doi:10.1038/nrg2487
Published online
25 November 2008

The process of the 'birth' of a new gene has fascinated biologists for a long time[1,2], not least because new genes are thought to contribute to the origin of adaptive evolutionary novelties and thus lineage- or species-specific phenotypic traits[1,3]. A major mechanism underlying the formation of new genes is gene duplication[2]. Traditionally, only DNA-mediated duplication mechanisms, that is, duplication of chromosomal segments containing genes, have been considered and widely studied in this context (reviewed in REFS 4,5). Nevertheless, gene copies originating through an alternative mechanism — the reverse transcription of mRNA intermediates — have been described since the early 1980s[6–8]. These intronless retroposed gene copies were long dismissed *a priori* as 'dead on arrival'[9–12], and routinely classified as processed pseudogenes[13] owing to the expected lack of regulatory elements and the presence of mutations, such as premature stop codons, in many copies. Indeed, they were mainly considered a nuisance and a confounding factor in transcription surveys because of their often high sequence similarity with parental genes.

However, after some anecdotal findings of functional retroposed genes in the late 1980s[14] an unexpectedly large number of functional retrogenes have recently been discovered, mainly in mammals and fruitflies[15–19]. These studies revealed that retrogenes have often evolved functional roles in the male germ line[16,17]. Other intriguing retrogene functions — for example, in antiviral defence[20], in hormone–pheromone metabolism[21,22], in the brain[23] or in courtship behaviours[24] — have also been postulated. More fundamentally, retrogene analyses have uncovered novel mechanisms for how new genes might arise (for example, the recruitment of regulatory elements) and obtain new functions (for example, through gene fusion and adaptive evolution). Finally, retroposed gene copies have served as unique genomic markers, increasing our understanding of various genomic processes ranging from the detection of extinct transcripts[25] to the origin of our sex chromosomes[17]. All of these findings were possible because of the growing number of complete genome sequences, and they were achieved by targeted cross-disciplinary approaches involving evolutionary analysis, mining of available large-scale expression data, and both molecular and genomics experiments.

This Review aims to cover the most exciting insights obtained from the study of RNA-based gene duplication, focusing on functionally relevant aspects of protein-coding retrogenes. Given that the process of retroposition (also known as retroduplication) has been most thoroughly studied and might be more frequent in mammals and fruitflies, we focus our discussion on these organisms. After a brief description of the process of retroposition, we discuss the abundance of retrocopies and functional retrogenes in mammals and *Drosophila* species. We then discuss how retrocopies might become transcribed and functional, and give an overview of novel mechanisms underlying the emergence of new gene functions that were uncovered in detailed surveys of young retrogenes. We then examine a major functional role of retrogenes in the male germ line, which is related to the biology and evolution of X chromosomes. Finally, we outline other general insights pertaining to mammalian genome evolution obtained from global retrocopy surveys, and conclude with potential future research directions.

Mechanisms of retroposition

To be heritable and hence of evolutionary relevance, retroposition needs to occur in the germ line. Thus, retroposition requires enzymatic machinery that not only can reverse transcribe and integrate fully processed cDNA copies of mRNAs from parental source genes into the genome, but that is also active in the germ line. The fact that retroposition relies on duplication through an mRNA intermediate also implies that only genes expressed in the germ line can be duplicated via this mechanism.

The key retroposition enzyme, reverse transcriptase, seems to stem from different types of retrotransposable elements, depending on the organism. In mammals, long interspersed nuclear elements (LINEs) seem to provide the enzymes necessary for retroposition. These retrotransposable elements encode a reverse transcriptase with endonucleolytic activity that can recognize any polyadenylated mRNA[26,27]. Esnault *et al.* and Wei *et al.* demonstrated that the L1 element subfamily of LINEs can generate processed genes[28,29], indicating that L1 retrotransposon activity has generated retroposed gene copies in mammals. The process of retroposition (including the hallmarks of retroposed gene copies) is detailed in FIG. 1.

Retrotransposable element-encoded enzymes are also likely to be responsible for retroposition in *Drosophila*[10,30] and some plants[31,32], which carry various retrotransposons with reverse-transcriptase activity. However, the retroposition machinery has not been studied in detail in these organisms to date. The paucity of retrocopies in non-mammalian vertebrates is probably explained by the lack of retrotransposons with reverse transcriptases that can process standard mRNAs. For example, bird genomes contain a relatively large number of CR1 LINEs[33], but CR1 reverse transcriptases cannot recognize polyadenylated mRNAs owing to their specificity towards a different target sequence, and are thus incapable of promoting retroposition of mRNAs from other genes[34]. The small number of RNA-based gene copies in birds[34] seems to have been mediated by retroviral mechanisms[35].

Rates of retrocopy and retrogene formation

Given that retrocopies are particularly abundant in mammals[11,17–19,36] owing to the high activity of L1 elements, we first discuss the rates of retrocopy and functional retrogene formation in mammals and then in *Drosophila* species. Thousands of retrocopies have been identified in several placental mammal (that is, eutherian) genomes[11,17,18,36]. This suggests a high rate of retrocopy formation during the evolution of this mammalian lineage. However, the rate of retroposition has not been constant, with periods of very high and low activity[11,37,38], which is probably due to the fluctuating activity of L1 elements (BOX 1). Recently, approximately 2,000 retrocopies were identified in the opossum genome[17], suggesting a similarly high retroposition rate in metatherians (that is, marsupials). Only ~80 retrocopies seem to be present in the platypus genome (H.K., N.V. and M.L., unpublished observations), which is consistent with the paucity of L1 elements in monotremes[39] — the most basal mammalian lineage.

It was long assumed that retroposed gene copies are mostly non-functional retropseudogenes because of their presumed lack of expression potential[10,13], although individual studies have revealed instances of functional retrogenes since the late 1980s[14]. But how many retrocopies have evolved into bona fide genes? Different types of evidence support the functionality of retrocopies; given the wealth of genomic data now available, the most straightforward approaches to look for retrogene functionality are based on evolutionary analyses that screen for signatures of selection. For example, the selective preservation of intact ORFs between distant species[17,18] or between several closely related species[37] can provide statistically significant and convincing evidence for non-neutral evolution of retrocopies — this therefore implies functionality. Furthermore, comparison of the rate of functionally relevant substitutions (that is, amino-acid changing) to the rate of neutral changes (that is, silent substitutions) in retrogene-coding regions can be used to detect non-neutral evolution, and is indicative of functional constraint[23,37].

In addition to such evolutionary approaches, molecular evidence can be used as an indication that a retrocopy is functional. One example is evidence of transcription, which can often be easily detected. Transcription alone is not sufficient to demonstrate functionality of individual genes, as non-functional DNA can be transcribed[18]. Evidence of translation (that is, the presence of a protein, which can be detected with specific antibodies), coupled with analysis of cellular phenotypes provides strong evidence of retrogene functionality. Ideally, the *in vivo* function of a retrogene is demonstrated — either by showing the association of retrogene mutations with disease[40–42], or by the targeted disruption of retrogenes in animal models[24,43,44]. However, given that solid experimental evidence for the functionality of retrocopies is currently hard to obtain on a larger scale, the estimates of overall rates of functional retrogene formation discussed in the following sections have largely been obtained from evolutionary and/or statistical analyses.

Vinckenbosch *et al.* estimated the number of functional retrogenes present in the human genome by comparing transcription levels of intact retrocopies with those of retropseudogenes, which reflect the transcriptional background noise in the genome[18]. They found that more than a thousand retrocopies show evidence of being transcribed[18], with intact retrocopies being transcribed to a much greater extent than retropseudogenes. On the basis of this observation the authors conservatively estimate that at least ~120 retrocopies are likely to be functional genes. Based on an assessment of selective constraint on primate retrocopies, Marques *et al.* estimated the rate of functional retrogene formation in primates[37]. They estimated that, on average, at least one functional retrogene per million years emerged on the primate lineage that led to humans[37].

In *Drosophila melanogaster*, in which the first retroposed gene copies were described in the early 1990s, a similar rate of functional retrogene formation was estimated[15,45]. Evidence of selective constraint suggests that about 90–100 functional retrogenes in this

Retroposition
A mechanism that creates duplicate gene copies in new genomic positions through the reverse transcription of mRNAs from source genes (also known as RNA-based duplication or retroduplication).

Retrocopy
Gene copy that results from the process of retroposition (also termed retroposed gene copy or retroposed copy).

L1 element
A member of the long interspersed nuclear element (LINE) family of repeats. Provides the enzymatic machinery necessary for the process of retroposition in mammals.

Retropseudogene
Non-functional retrocopy, which usually carries frameshift-causing insertions or deletions and/or premature stop codons that preclude gene function.

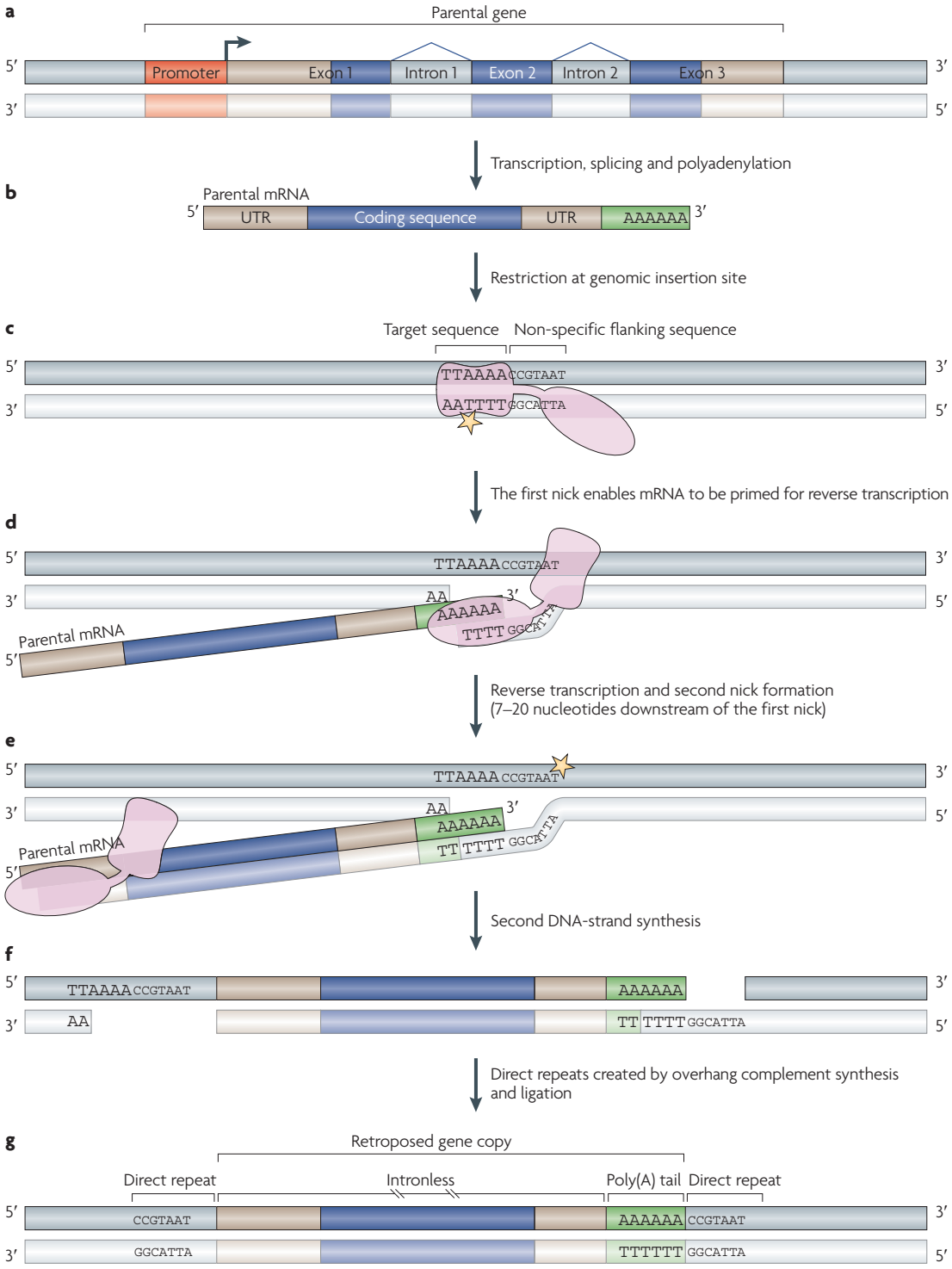

Figure 1 | **Mechanism of gene retroposition. a** | Gene retroposition is initiated with the transcription of a parental gene by RNA polymerase II. **b** | Further processing of the resulting RNA (by splicing and polyadenylation) produces a mature mRNA. **c** | Gene retroposition is mediated by the L1 endonuclease domain (pink rectangle), which creates a first nick (yellow star) at the genomic site of insertion at the TTAAAA target sequence. **d** | This nick enables the mRNA to be primed for reverse transcription by the L1 reverse transcriptase domain (pink oval), which uses the parental mRNA as a template. **e** | Second-strand nick generation (precise mechanism not known). **f** | Second DNA-strand synthesis (precise mechanism not known). **g** | cDNA synthesis in the overhang regions created by the two nicks. This process creates a duplication of the sequence flanking the target sequence, which is one of the molecular signatures of gene retroposition; other signatures include the lack of introns and the presence of a poly(A) tail. The direct repeats and the poly(A) tail degenerate over time, and are therefore usually only detectable in recent retrocopies. The illustration is based on findings described in REFS 26–28.

Box 1 | **Retrocopies as genomic archives**

Retrocopies can serve as useful genomic markers of transcript activity during evolution. For example, because retroposition is mediated by long interspersed nuclear elements (LINEs) the rate of retrocopy generation, which might be calculated on the basis of the divergence of retrocopies and parental genes at a synonymous site, can be used to explore the activity of LINE retrotransposons during evolution.

Moreover, given that the probability of retroposition of a gene is expected to mainly depend on the abundance of its transcripts in the germ line, the number of retrocopies should reflect parental gene activity during these stages[11,12]. Consistently, well-known housekeeping genes and genes with high germline expression levels have produced many retrocopies[11,12,101]. Thus, retrocopies might serve as unique markers to shed light on the tissue origin of retroposition by correlating parental gene expression during different male and female germline stages with the abundance of their retrocopy offspring in the genome. The better the correlation observed in such an analysis, the higher the number of retrocopies that would have emerged in a given germ line or embryonic cell type.

Finally, the fact that retrocopies reflect their parental transcript structures has been exploited to detect previously unannotated or extinct 'fossil' transcripts[25,102]. For example, in a recent study, the authors reconstructed ancestral transcripts that were present in the common ancestor of humans and chimpanzees, using retrocopy sequences and inferred potential exon gains and losses in humans and chimpanzees based on their analysis[102].

invertebrate lineage are functional[15,16,46]. However, the total number of retrocopies in the *Drosophila* genus is much lower than that in mammals. This seems to be due mainly to the paucity of retropseudogenes in *Drosophila* genomes[9,47] (owing to the extremely short half-life of unconstrained DNA in this genus[9]) rather than to a low rate of retroposition.

Sources of regulatory elements

The observation that a significant number of retrocopies have evolved into bona fide genes raises the question of how retrocopies can be expressed in their new genomic location. To become expressed at a significant level and in a meaningful way (for example, in tissues in which it can exert a selectively beneficial function), a new gene needs to obtain a core promoter and probably other elements, such as enhancers, that regulate its expression. In this section, we discuss various mechanisms through which the acquisition of promoters and other regulatory elements might occur.

Generally, the expression of retrocopies might benefit from the presence of pre-existing regulatory elements in their vicinity. For example, a straightforward way for a retrocopy to obtain transcription potential would be to 'hitch-hike' on the regulatory machinery of other genes. Indeed, a number of cases have been described in which retrocopies are located in an intron of a host gene, and are transcribed in the form of a fusion transcript together with host gene exons[18,41,48,49] (FIG. 2a). In mammals, retrocopies are often transcribed together with only 5′-UTR exons of the host gene, as splice variants, thus potentially avoiding interference with host gene functions[18]. In general, transcribed retrocopies tend to be close to other genes, suggesting that their transcription might be facilitated by the open chromatin and/or regulatory elements of nearby genes[18] (FIG. 2b). This possibility is supported by observations that retrogenes might be transcribed from bi-directional CpG-rich promoters of genes in their proximity (H.K., unpublished observations). The sometimes substantial distances between the retrocopy insertion site and these promoters are usually spanned by new 5′ untranslated exon–intron structures that arose during the process of promoter acquisition[18].

In a similar way, that is, via the acquisition of new 5′-UTR structures, retrocopies might also become transcribed from distant CpG-enriched sequences, which often have inherent capacity to promote transcription[50], and that are not previously associated with other genes (H.K., unpublished observations) (FIG. 2c). These distant CpG 'proto-promoter' elements might have been optimized by natural selection after they became associated with a functional retrogene. Similarly, distant promoters from retrotransposable elements might have been 'captured' by retrocopies for their transcription via the acquisition of new 5′ untranslated exon–intron structures. In addition, retrotransposons[51] (or, potentially, CpG-island proto-promoters) that are immediately upstream of retrogene insertion sites might also be used directly (FIG. 2d).

Until recently, it was thought that retrocopies are unlikely to directly inherit parental promoters (hence the common expectation that they are unlikely to evolve into functional genes), although instances of parental-promoter inheritance had been found[52–54]. However, a recent study suggests that retrocopies could frequently inherit basic promoters directly from their parental source genes[55]. Often, these parental genes are transcribed from CpG promoters, which usually have multiple transcriptional start sites[56] (TSSs). If a retrocopy stemmed from a parental transcript with a TSS located far upstream, the mRNA that gave rise to the retrocopy might carry downstream promoter sequences and TSSs with sufficient capacity to promote transcription (FIG. 2e). The frequent inheritance of CpG promoters might also help to explain why a significant number of retrogenes evolved paternally or maternally imprinted expression[57,58] (TABLE 1).

In *Drosophila* spp., the source of transcription potential of retrogenes is somewhat more elusive. Although, similarly to mammals, host gene fusions have occurred in this genus[48,49] and retrogene transcription might be facilitated through the transcriptional activity of genes in their vicinity[15], some other mechanisms described for mammals, such as parental promoter inheritance or retrotransposon-driven transcription, have not yet been detected in fruitflies. Instead, small substitutional changes in pre-existing upstream sequences of retrogene insertion sites that occurred under the influence of natural selection have been postulated to play a part in the formation of basic *Drosophila* retrogene promoters[15,59] (FIG. 2f).

We note that the various mechanisms described here that might endow retrogenes with regulatory elements probably often only provide the basic means for the initial transcription of retrocopies, whereas more sophisticated regulatory elements might evolve with time (see for example, the mammalian phosphoglycerate kinase 2 (*Pgk2*) retrogene[52,60]) (TABLE 1).

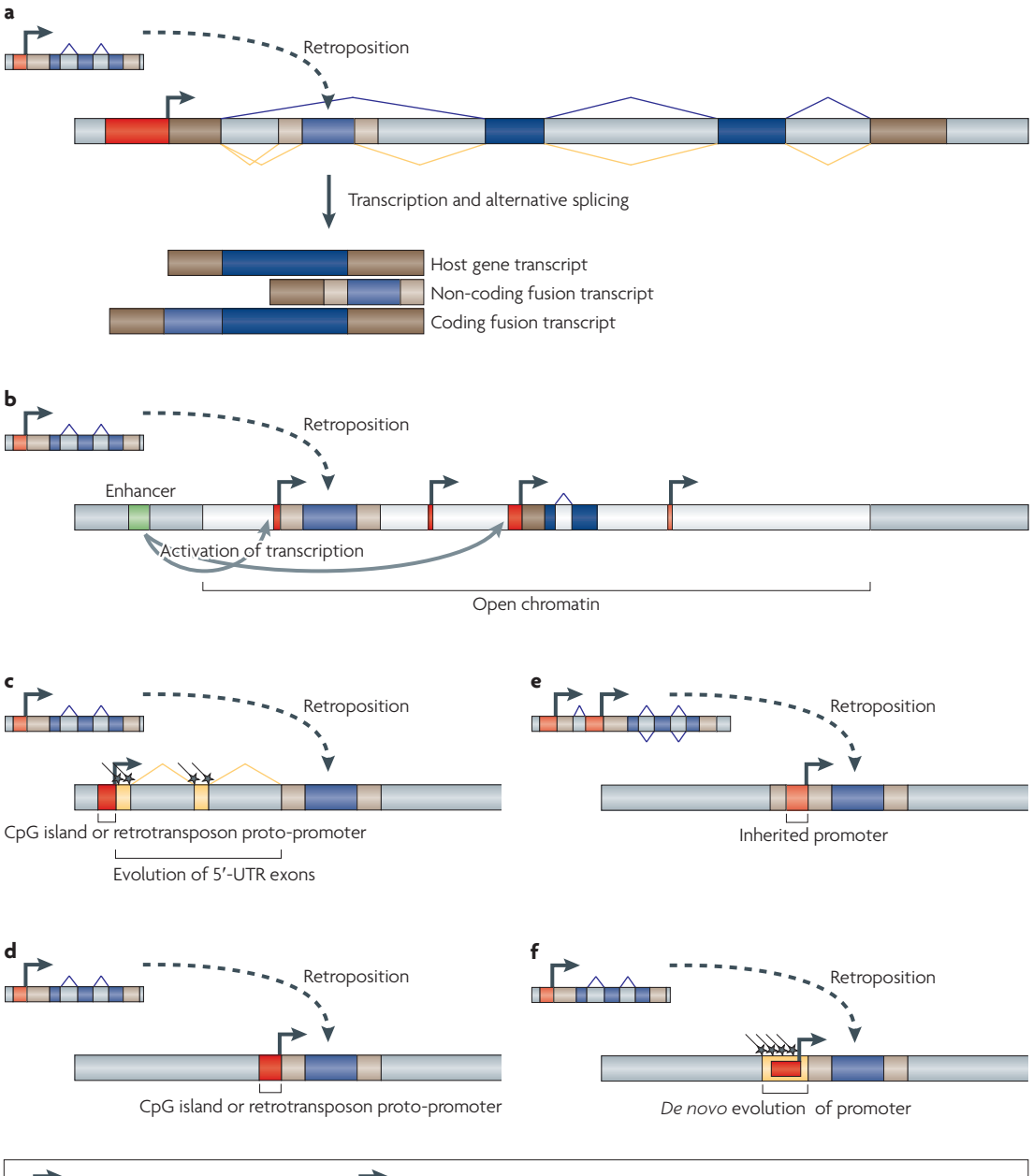

Figure 2 | **Source of retrogene promoters.** This figure illustrates various scenarios that lead to the transcription of retroposed gene copies. **a** | Retrocopies can insert into intronic sequences of host genes. The evolution or presence of splicing signals enables these copies to be integrated into new splice variants of their host gene. Depending on the localization of these new splice sites, these variants result in either non-coding fusion transcripts (the entire ORF derives from the retrocopy) or coding sequence fusions (the coding region of the retrocopy is fused to that of the host gene). **b** | The insertion of retrocopies into actively transcribed regions with an open chromatin structure facilitates their transcription, as this increases accessibility for the transcriptional machinery. The presence of enhancer elements from neighbouring genes and weak transcription promoting sequences (not previously associated with genes) can further strengthen their transcriptional activity. **c** | Recruitment of distant promoters in the genomic neighbourhood via the acquisition of a new untranslated exon–intron structure. **d** | Recruitment of proto-promoters from retrotransposons or CpG islands. **e** | Inheritance of parental promoters through alternative transcriptional start site use by the parental gene. **f** | *De novo* promoter evolution in the 5′ flanking region of the insertion site by single nucleotide substitutions.

The evolution of new functions from retrogenes

DNA versus RNA-based duplication. The fundamental differences between the two major duplication mechanisms — segmental duplication (reviewed in REFS 4,5) and retroposition — have significant consequences for the evolutionary fates of resulting gene copies and their analysis. Segmental duplication regularly produces daughter copies that inherit the genetic features — exons, introns and regulatory elements — of the ancestral gene, whereas retroposed copies usually lack introns and are less likely to have strong regulatory elements following their emergence. Therefore, segmental duplication is more likely to yield expressed daughter copies than the retroposition process. In addition, segmental duplicates are likely to exhibit similar expression patterns in their early evolution, which can often imply that one copy is

Table 1 | **Representative retrogenes in mammals and fruitflies**

Genes	Phylogenetic distribution	Features (chromosomal origin, structure, type of selection or function)	Refs
Primates			
GLUD2	Hominoids	Into the X, positive selection, subcellular adaptation, adaptation to neurotransmitter glutamate metabolism	23,67
CDC14Bretro	Hominoids	Positive selection, subcellular adaptation, derived from cell-cycle gene, brain- and testis-specific expression	37,65
c1orf37-dup	Humans	Positive selection, transmembrane protein	66
PGAM3	Old World primates	Positive selection, phosphoglycerate mutase	64
TRIM5–CypA	Macaque lineage	Chimeric gene, retrovirus restriction, *CypA* portion derives from retroposition	72–74
TRIM5–CypA	New World monkeys	Chimeric gene, retrovirus restriction, *CypA* portion derives from retroposition	20
PIP5K1A–PSMD4	Hominoids	Chimeric gene, positive selection, subcellular change, fusion retrogene — stems from chimeric transcript of two adjacent parental genes	75
TAF1L, KIF4B	Old World primates	X-derived	37,103
RBMXL1	Old World primates	X-derived, chimeric gene, fusion to host gene UTR	37
Utp14c	Primates	X-derived, chimeric gene, evidence that it is required for male fertility, fusion to host gene UTR	40
Rodents			
Utp14b	Rodents	X-derived, chimeric gene, required for male fertility, fusion to host gene UTR exon	41,42
U2af1-rs1	Rodents	X-derived, paternally imprinted	57
PMSE2b	Mouse*	Inserted into a LINE1 that drives its transcription	51
Mammals			
Cstf2t	Eutherians	X-derived, chimeric gene, required for male fertility, crucial for proper polyadenylation in meiosis and post-meiosis	43
HNRNPGT	Eutherians	X-derived, required for male fertility	44
Pgk2	Eutherians	X-derived, promoter inherited from parent, acquisition of a testis-specific enhancer, first described X-derived retrogene	14,60
Inpp5f, Nap1/5, Mcts2	Eutherians	X-derived, paternally imprinted, located in introns of host genes	57
KLF14	Eutherians	Maternally imprinted, accelerated evolution on the human lineage	58
USP26	Eutherians	Into the X, among the five most positively selected genes in human–chimp comparison	104
Drosophila			
jingwei	*Drosophila yakuba*, *Drosophila santomea* and *Drosophila teisseri*	Chimeric gene, positive selection, retrocopy encoded ADH domain evolved new substrate (alcohol) specificity	21,48
sphinx	*Drosophila melanogaster*	Chimeric gene, positive selection, retrocopy evolved into non-coding RNA gene that promotes male–female courtship	24,49
Adh–Twain	*Drosophila subobscura*, *Drosophila guanche* and *Drosophila madeirensis*	Chimeric gene, positive selection, putative functional adaptation to new substrate specificity	105
mojoless	*Drosophila* genus	X-derived, required for male fertility	106
Dntf-2r	*D. melanogaster* subgroup	Substitutions in an upstream proto-promoter element seem to have provided this gene with a new, testis-specific promoter	59

The cases listed here are representative of the different mechanisms that lead to the formation of retrogenes, their chromosomal distribution and the type of function they can obtain. We describe most of these genes in the main text. *Identified in the mouse, phylogenetic distribution not established. ADH, alcohol dehydrogenase; LINE1, long interspersed nuclear element 1.

initially functionally redundant. The increased gene dose might even be deleterious, although increased gene dosage can sometimes be beneficial and thus selectively preserved. By contrast, retroposed copies often need to recruit regulatory elements to become transcribed (see above). However, this also means that retrocopies that do become transcribed are probably more prone to evolve new expression patterns and, as a consequence, novel functional roles than gene copies that arise from segmental duplication.

A further fundamental difference between the two duplication mechanisms is related to the relationship between the two duplicate members of the pair. The clear directionality in the retroposition process, which is often not discernible for segmental duplications, facilitates studies on the origin of new gene functions. This is because parental genes usually maintain the ancestral gene function, although there are interesting exceptions to this rule[61], whereas new functions usually are acquired by the intronless daughter retrogene copies. This directionality also renders the detection and analysis of young duplication events straightforward; these duplication events are particularly informative for the study of new gene functions (see below). However, recent segmental duplicates are not easily distinguishable and are more difficult to study because they are, for example, frequently collapsed into a single locus in standard genome assemblies owing to their high sequence and structural similarities.

Finally, retroposition usually produces gene copies on chromosomes different from that of the parental gene copy, whereas segmental duplications are less likely to involve different chromosomes — although the rate of inter versus intrachromosomal segmental duplication differs between lineages[45,62,63]. Thus, retroposition is the ideal 'vehicle' for interchromosomal gene 'movements', the directions of which are also easily determined based on the inherent directionality of the process (see below for a detailed discussion of retrogene movement studies).

Nevertheless, owing to the abundance of functional segmental duplicates in nearly all genomes studied, numerous studies of segmental duplication have yielded many fundamental insights and have established general concepts regarding the emergence of new gene functions (reviewed in REFS 4,5).

However, because of the particular features of retroposed gene copies outlined above, the analysis of retroposition has provided additional insights with respect to the functional evolution of new genes not previously described for segmental duplicates. In particular, the analysis of young retrogenes has provided novel insights into mechanisms underlying the evolution of new genes, as the changes in sequence that occurred during their early evolution are usually still traceable using evolutionary approaches[1]. In mammals, the study of young retrogenes has mainly focused on primate cases. Systematic surveys and individual studies led to the discovery of several young retrogenes that emerged on the primate lineage leading to humans[23,37,64–66]. For some of these, positively selected substitutions could be tied to functional change and adaptation[23,65,67] (TABLE 1).

Subcellular adaptation
A process by which a duplicate gene product evolves a new localization in the cell or localizes more specifically to one of the ancestral compartments under the influence of positive Darwinian selection.

Emergence of new cell compartment-specific functions. Further analysis of these recently emerged retrogenes uncovered a novel mechanism underlying the emergence of new gene function. They showed that new gene functions can arise through changes in the localization of encoded proteins in the cell, a process that is termed subcellular adaptation[65,67,68]. The following examples demonstrate two ways by which this process might occur (FIG. 3).

The glutamate dehydrogenase 2 (*GLUD2*) retrogene exemplifies one form of subcellular adaptation called sublocalization[68], in which the protein encoded by the new gene becomes more specifically targeted to one or several of the ancestral cellular compartments. *GLUD2* (TABLE 1) emerged in the common ancestor of humans and apes 18–25 million years ago by retroposition from its parental gene, *GLUD1*, which encodes an enzyme that degrades glutamate[69]. The enzyme encoded by *GLUD2* evolved unique biochemical properties soon after the duplication event through two key amino-acid substitutions that were fixed as a result of positive selection[23]. These changes were suggested to reflect the functional adaptation of GLUD2 to the metabolism of the neurotransmitter glutamate in the brain[70]. A further study of *GLUD2* uncovered another level of functional adaptation. Rosso *et al.* showed that whereas the ancestral glutamate dehydrogenase enzyme localizes to mitochondria and the cytoplasm, GLUD2 became specifically targeted only to the mitochondria, owing to a single, positively selected substitution in its N-terminal targeting sequence[67]. This event probably contributed to the adaptation of GLUD2 to a function in glutamate metabolism in the brain and other tissues. Thus, GLUD2 is an example of rapid change in subcellular localization and function of a new protein that has been driven by natural selection[65,67,68] (FIG. 3a).

The analysis of another ape-specific retrogene, *CDC14Bretro*, revealed that proteins encoded by new genes can completely relocalize to new, previously unoccupied cellular niches during evolution under the influence of natural selection. This process is a variant form of subcellular adaptation termed subcellular relocalization, or neolocalization[68,71]. *CDC14Bretro* stems from a splice variant of the *CDC14B* cell-cycle gene[65] (TABLE 1) and it encodes a protein that became specifically expressed in the adult and fetal brain and testes soon after its emergence in the common human and ape ancestor. It then completely relocalized in the cell owing to intense positive selection in the common African ape ancestor ~7–12 million years ago, shifting from the ancestral role of stabilizing microtubules to a localization and function in the endoplasmic reticulum (FIG. 3b).

Notably, a recent global survey of yeast duplicate proteins, which was prompted by these retrogene studies, showed that subcellular adaptation seems to be widespread, and is involved in the evolutionary fate of at least 30% of duplicates[68]. Thus, in conclusion, the analysis of young retrogenes led to the finding that, in addition to changes in gene expression and/or the biochemical function of the protein through neofunctionalization or subfunctionalization[5], rapid and selectively

Domain shuffling
Juxtaposition of one or more exons from two different genes that encode functional protein domains.

driven subcellular adaptation by either neolocalization (*CDC14Bretro*) or sublocalization (*GLUD2*) is a common, previously underappreciated mechanism underlying the emergence of new gene function (FIG. 3).

Gene fusion and domain shuffling. New gene functions can also arise through gene fusion, which is defined as the fusion of two previously separate source genes into a single transcription unit[1]. Gene fusion might occur through various mechanisms, including DNA-based recombination events, and can lead to the juxtaposition of exons encoding functional protein domains from different genes, in which case it is a form of exon or domain shuffling[1].

Fusion of retroposed gene copies to the genes in which they have inserted has yielded new genes with important functions. Detailed studies of such fusion genes uncovered surprising aspects of new-gene formation, such as a recurrence of the fusion of genes with complementary functions in the case of the *TRIM5–CypA* fusion gene (FIG. 4). A retroposed copy of the *CypA* gene, which encodes a protein that potently binds retroviral capsids, was shown to have integrated independently into the antiviral defence gene *TRIM5* in a New World monkey[20] (FIG. 4a) and in an Old World monkey[72–74] (FIG. 4b). In both cases, the retrocopy-encoded CypA protein replaced and functionally substituted the original capsid-binding domain (B30.2) from TRIM5. The TRIM5–CypA fusion protein more efficiently restricts HIV-1 and other retroviruses than the ancestral TRIM5 (REFS 20,72–74). The *TRIM5–CypA* gene fusion is a striking case of domain shuffling and convergent evolution. The seemingly unlikely multiple independent insertions of *CypA* retrocopies into the same gene in different species were probably facilitated by the high retroposition rate of the *CypA* gene, which is due to its high expression in the germ line. Rare *TRIM5–CypA* fusions were then probably driven to fixation during the evolution of the monkey lineages by strong selective pressures, because potent TRIM5 variants can provide a high degree of resistance to lethal and common diseases caused by various retroviruses[73].

Recent studies revealed that fusion genes can also arise through the co-retroposition of adjacent parental source genes. Akiva *et al.* identified a recently retroposed gene (*PIPSL*) on human chromosome 10 that stems from a fusion transcript of two parental genes (*PIP5K1A* and *PSMD4*) that are next to each other on chromosome 1 (REF. 75). Babushok *et al.* then showed that the gene was exclusively expressed in testes in humans and chimpanzees[76]. But curiously, although *PIPSL* was apparently shaped by strong positive selection — suggesting functionality and adaptive evolution of the encoded protein — this fusion gene seemed to be post-transcriptionally repressed. However, in a recent follow-up analysis, evolutionary and experimental support was obtained for the functionality of this gene in hominoids (M.L., unpublished observations). Given the abundance of intergenic splicing in mammals[75,77], we speculate that co-retroposition of adjacent genes might potentially be responsible for the origination of other chimeric retrogenes.

Analysis of chimeric genes in *Drosophila* species has demonstrated how gene fusion via retroposition can generate raw material for the evolution of new gene functions under the influence of positive Darwinian selection. The gene *jingwei* (*jgw*), which was the first chimeric gene involving retroposition described in any species[48], originated by the insertion of a retrocopy of

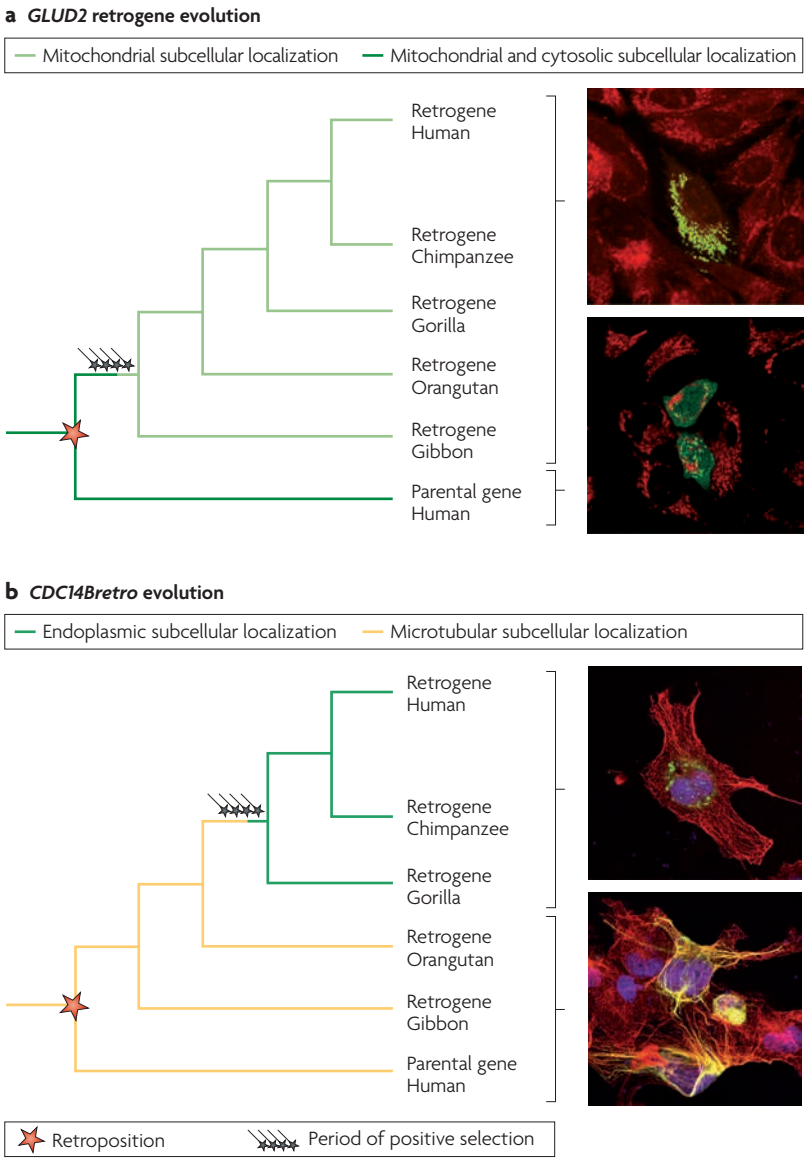

Figure 3 | **Subcellular adaptation of proteins encoded by new duplicate genes.** The adaptive evolution of two primate-specific retrogenes: glutamate dehydrogenase 2 (*GLUD2*) (**a**), and *CDC14Bretro*, which stems from a splice variant of the *CDC14B* cell-cycle gene (**b**). Phylogenetic trees indicate retroposition events; periods of adaptive evolution and reconstructed subcellular localizations are indicated. Microscopy images display representative subcellular phenotypes for the indicated branches. For the *GLUD2* images (**a**), protein localization is in green and mitochondria are red. For the *CDC14Bretro* images (**b**), protein localization is in green, nuclear DNA is blue, and microtubules are red. Yellow signals indicate an overlap of the protein with microtubules. The microscopy images for part **a** are reproduced from REF. 67. The microscopy images for part **b** are reproduced from REF. 65.

Meiotic sex chromosome inactivation
(MSCI). The transcriptional silencing of the X and Y chromosomes during the meiotic phase of spermatogenesis.

the *Alcohol dehydrogenase* gene (*Adh*) into the *yande* gene[48] (TABLE 1). The functional evolution of *jgw* was recently analysed using a biochemical approach[21,22], which revealed that the JGW protein (particularly the ADH domain) was shaped by positive selection and apparently evolved a role in hormone and pheromone biosynthesis or degradation processes.

The *Drosophila sphinx* (*spx*) gene[49] (TABLE 1) illustrates a mechanism for how RNA genes with important new functions can arise, a process that is currently poorly understood. The *sphinx* gene emerged within the last 2–3 million years and derives from a retroposed ATP synthase gene that fused to exons located in the vicinity of the insertion site. Notably, the retroposed gene copy lost its protein-coding capacity by accumulating nonsense mutations, and *spx* subsequently evolved into a non-coding RNA gene under the influence of positive selection. Dai *et al.* knocked out the *spx* gene in *D. melanogaster*[24], which caused an increase in male–male courtship behaviour relative to wild-type flies, suggesting that *spx* is the first recently emerged gene for which a behavioural phenotype has been identified.

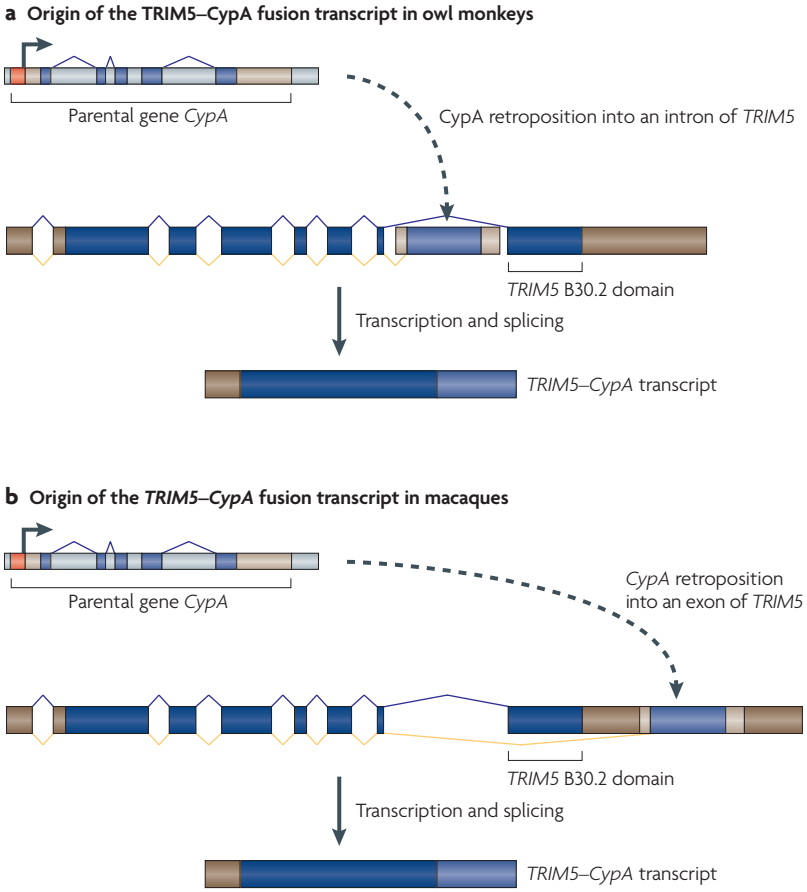

Figure 4 | **Origin of *TRIM5–CypA* gene fusions in owl monkeys and macaques.** **a** | Retroposition of *CypA* (encoding a protein that binds retroviral capsids) into an intron of the antiviral defence gene *TRIM5* from owl monkeys. The resulting fusion gene is shown (similar to the process displayed in FIG. 2). **b** | An independent retroposition of *CypA* into the UTR of *TRIM5* in macaques is shown, also resulting in a new *TRIM5–CypA* fusion gene.

Testis functions and sex chromosome evolution

Global surveys of retroposition in mammals and fruitflies have shown that retrogenes have often evolved functions in the testes. The formation and preservation of many of these genes is closely linked to the biology and selective forces, imposed by the male germ line, that have shaped X chromosomes since their emergence. These issues, and how dating of the origin of these retrogenes has also allowed a reassessment of the age of mammalian sex chromosomes, are discussed below.

Expression in testes. Numerous retrogene studies in both mammals and fruitflies revealed an overall propensity of retrogenes to be expressed in the testis[16,18,37,46,48]. A combination of a testis-expression bias and natural selection was postulated to explain this observation[17,37]. In meiotic and post-meiotic spermatogenic cells, the autosomal chromosomes seem to be in a state of hypertranscription owing to various modifications of the chromatin (reviewed in REF. 78). It was suggested that this hypertranscription state enables transcription of DNA that is usually not transcribed. It might therefore have facilitated transcription of retrocopies[37] but also other types of duplicates[79] in the testis during their early evolution. A subset of these retrocopies subsequently obtained beneficial functions in the testis and evolved into bona fide genes (see below). Natural selection then further enhanced their promoters and other regulatory elements, which led to a stronger and more refined testis-expression pattern among the functional retrogenes.

An alternative and not mutually exclusive hypothesis is based on the notion that retrocopies might preferentially insert into open, actively transcribed chromatin[80]. Given that retroposition occurs in the germ line, retrocopies might predominantly insert into or close to germ line-expressed genes, which would facilitate retrocopy transcription in the germ line. However, in *Drosophila* species, this hypothesis seems to explain testis expression of only some retrogenes[81]. In mammals, this insertion-bias scenario remains to be explored.

Retrogenes 'out of the X'. As noted above, the retroposition process readily produces gene copies on chromosomes different from that of the parental gene copy. Global genomic surveys of such gene 'movements' revealed an intriguing pattern that was observed both in mammals[17–19] and in *Drosophila*[16]: a disproportionately large number of parental genes on the X chromosome have given rise to functional retrogene copies on autosomes[16,19] (FIG. 5a). For mammals, it was shown that these autosomal retrogenes are specifically expressed in the testis during and after the meiotic stages of spermatogenesis, whereas their X-linked parents (usually broadly expressed housekeeping genes) are transcriptionally silenced during these stages owing to male meiotic sex chromosome inactivation (MSCI)[17] (reviewed in REF. 82) (FIG. 5a).

Importantly, these mammalian X-derived retrogenes are significantly more frequently and more specifically expressed during and after meiosis than other retrogenes[17], which also tend to be expressed in testes (see above). This substantiates the hypothesis that retrogenes

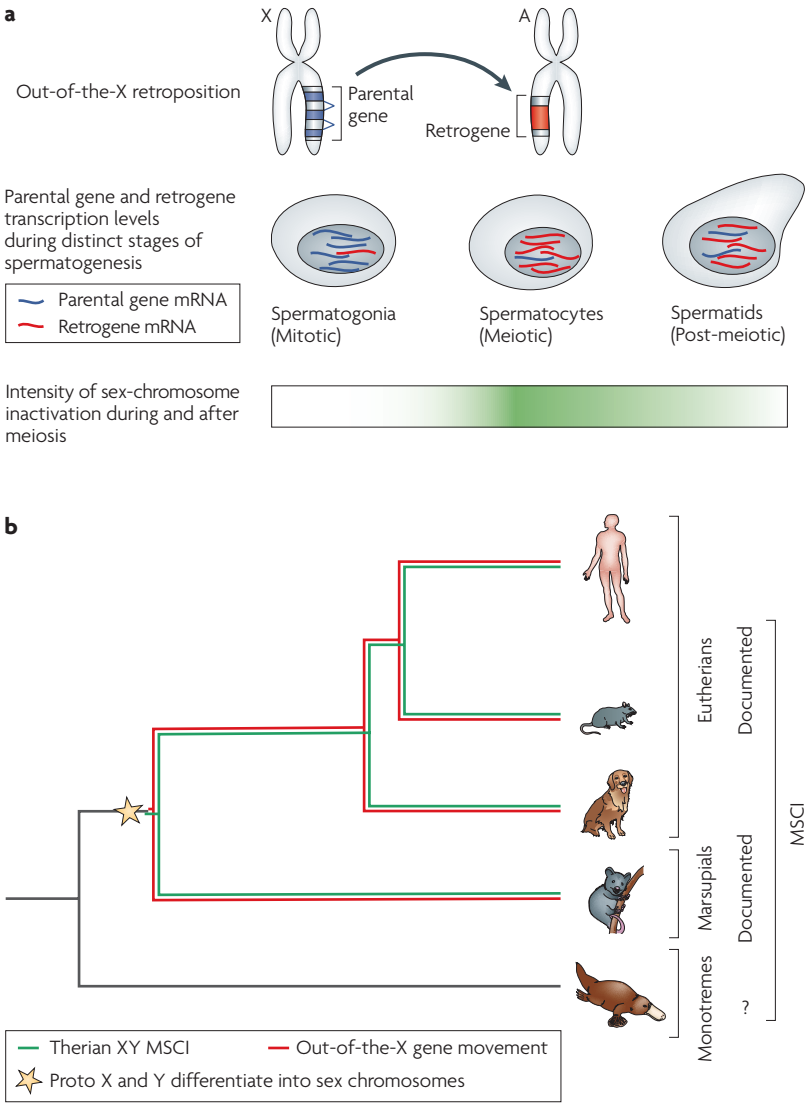

Figure 5 | **Retrogenes, meiotic sex chromosome inactivation (MSCI) and the emergence of mammalian sex chromosomes. a** | Illustration of the retroposition of an X-linked parental gene to an autosome (top). Illustration of the expression of X-linked parental genes and their autosomal retrogene copies before (spermatogonia), during (spermatocytes) and after (spermatids) the process of MSCI (bottom). **b** | The evolutionary onset for the selectively driven out-of-the-X retroposition process and MSCI, as well as the inferred origin of therian (placental mammals and marsupials) sex chromosomes.

that stem from the X chromosome have been fixed during evolution and shaped by natural selection to compensate for parental, housekeeping gene silencing during and after MSCI[17,19,83]. This compensation hypothesis has also been supported by functional studies that showed that loss of function of retrogenes with X-linked progenitors lead to severe defects of male meiotic functions in mice[41–44] and probably in humans[40]. Curiously, the potential mechanistic biases favouring expression in meiotic and post-meiotic cells (see above) allow X-derived retrogenes to be expressed precisely where needed to compensate for the lack of expression from the parental gene. Thus, together with the fact that the retroposition process readily moves genes between chromosomes, this means that retrogenes — rather than DNA-based duplicates — might easily evolve into functional autosomal substitutes of their X-linked parental genes during the late stages of spermatogenesis.

Although it was recently suggested that the major cause for the out-of-the-X movement in *Drosophila* species might be different from that in mammals[84], a recent study suggests that MSCI occurs in *Drosophila*[85]. Therefore, MSCI might be the main force responsible for the preferential fixation of X-derived retrogenes with meiotic or post-meiotic expression in fruitflies as well as in mammals. In addition, similarly to mammals, retrogene–parental gene expression patterns also seem to be complementary during meiosis in *Drosophila*[46].

The origin of mammalian sex chromosomes. A recent survey of young retrogenes in primates showed that the out-of-the-X movement of retrogenes is ongoing[37], which suggests that gene export from the X chromosome continues to be selectively beneficial. But when did this process begin during evolution? A systematic dating analysis using representative genomes from the three major mammalian lineages recently revealed that, although retrogenes have been generated since the common ancestor of all mammals, selectively driven retrogene export from the X chromosome only started later, in the eutherian and marsupial lineages[17] (FIG. 5b). Given that MSCI is the probable selective force that is driving genes off the X chromosome, this observation suggested that MSCI emerged, rather late, in the common ancestor of eutherians and marsupials, well after their separation from the monotreme lineage[17] (FIG. 5b).

Moreover, these observations have led to a reassessment of the age of our sex chromosomes, which evolved from an ancestral pair of autosomes[86,87]. Given that MSCI probably reflects the spread of the recombination barrier between the X and Y chromosomes during their evolution[17,88], Potrzebowski *et al.* concluded that these chromosomes originated, probably at a late stage, in the common ancestor of eutherians and marsupials and not in the common ancestor of all mammals, and are therefore much younger than previously thought[17] (FIG. 5b). This view is supported by the recent analysis of the platypus genome, which revealed that monotreme sex chromosomes share homology only with bird and not with therian (placental mammals and marsupials) sex chromosomes[39,89,90].

Retroposition 'into the X'. Curiously, retrogenes are not only exported from the X chromosome, but they are also preferentially imported into this chromosome in mammals[19]. There seems to be a slight mechanistic bias that favours the insertion and/or retention of retrocopies on the X chromosome[19]. Although the cause of this bias remains unclear, the excess of retropseudogenes on the X chromosome is consistent with the accumulation of other non-functional retro-elements (including LINEs) on the X chromosome in the mammalian lineage[91]. However, a strong selective force — the precise nature of which remains to be identified — has apparently led to the preferential fixation of bona fide retrogenes on the

X chromosome[19]. Finally, note that no increased fixation rate of retrogenes on the X chromosome is observed in *Drosophila* species[16,92]. This might reflect differences in the biology of sex chromosomes between mammals and fruitflies, but the precise reasons for this discrepancy need to be clarified.

Retroposition and gene structure evolution

Studies of the process of retroposition have not only shed light on the origin of new genes as discussed above, but have also provided other general insights pertaining to the evolution of mammalian genomes. These findings are discussed below and in BOX 1, which highlights how retrocopies reflect aspects of transcriptome evolution.

Retroposition and intron loss. One way by which retroposition has shaped mammalian genes is by mediating the loss of introns. Intron gains are rare events during evolution, whereas intron loss seems to be more frequent[93]. In mammals, for example, not a single case of intron gain has been documented, whereas more than 100 intron losses have been reported[94]. These intron losses seem to have been mediated by recombination of the gene displaying intron loss with the reverse transcribed, processed mRNA molecule (the cDNA) of the same gene[94,95]. The lines of evidence that support this hypothesis include: the always precise loss of the intronic sequence — the alternative mechanism, DNA deletion, would often result in imprecise intron loss; the fact that intron loss usually affects genes that are highly expressed in the germ line, thus producing many processed cDNAs that might recombine with the source gene; and the preferential loss of introns towards the 3′ end of the genes[94,96], reflecting that reverse transcription begins at the 3′ end of transcripts — thus, incomplete 3′ cDNAs can recombine with the source gene, leading to 3′ intron loss.

Retrogenes and splicing constraints. Retrogenes have also helped to support the novel hypothesis that the preservation of splicing signals constrains protein evolution. Specifically, a recent study suggested that the selective pressures on splice signals (enhancers and silencers) near exon boundaries significantly reduce the rate of protein evolution[97]. The rate of protein evolution of retrogenes is highest near the sequences in which intron–exon junctions previously resided in the parental genes that gave rise to the retrogenes. Therefore, splicing sequence constraints might have hampered the evolution of multi-exon gene-encoded proteins, thus potentially preventing functional optimization of proteins. It will be interesting to test whether retrogenes have evolved more efficient and/or adapted proteins compared with their intron-containing parents, owing to the relaxation of splicing constraints.

Conclusions

mRNA-derived duplicates were long thought to be doomed to pseudogenization and decay. However, a significant number of retroposed gene copies have escaped this evolutionary fate and have evolved into bona fide genes, as outlined in this Review. Retroposed genes are probably still much less likely to become functional compared with 'normal' DNA duplicates owing to their peculiar properties, which include the frequent lack of strong regulatory elements following their emergence. However, because of these properties retrogenes often evolved in unique ways, being much more prone to evolve new expression patterns, new genomic locations and new functions compared with DNA duplicates. Thus, individual and global surveys of retrogenes, using a variety of evolutionary, genomic and molecular tools, have unearthed previously unknown molecular mechanisms pertaining to the origin of new genes; for example, promoter recruitment and subcellular adaptation of encoded proteins. These surveys have also provided unexpected and unique insights into genome evolution; for example, the origin and evolution of our sex chromosomes.

In spite of these recent advances in the RNA-based duplication field, much remains to be done. So far, only a few young retrogenes have been pinpointed, and even fewer studies (most of which are discussed in this Review) have attempted to characterize the functional evolution of young retrogenes, thus going beyond mere descriptions of evolutionary signatures. Future work should therefore first aim to identify more young functional retrogenes. Such studies are challenging (at least in mammals) owing to the difficulty in assessing their selective preservation, but they will benefit from the steadily increasing number of available complete genome sequences in primates. Notably, very recent functional hominoid retrocopies might soon be identified because of the astounding number of human genomes that will be completed using the recently developed ultra-high-throughput sequencing technologies[98]. New cases of young retrogenes should then be subjected to in-depth analyses of their functional evolution, using evolutionary analysis combined with molecular, cellular and *in vivo* experiments; for example, transgenic mice carrying primate-specific genes, or knockout studies in *Drosophila*. Ultimately, such studies are likely to uncover additional modes underlying the evolution of new gene function and provide a more global view of the contribution of retrogenes to cellular or organismal phenotypes.

It will also be interesting to screen for retrogenes in other organisms for which complete genomes are becoming available and to study their chromosomal localization patterns, evolution and functions. For example, a recent study discovered a surprisingly large number of functional retrogenes with interesting properties in the rice genome[32], a large proportion of which were fused to other genes. This large number of retrogenes was unexpected, given that the retroposition activity in plants was traditionally thought to be low.

We believe that retrocopies generally are still a relatively untapped resource and are likely to reveal further unpredicted and fascinating aspects, which might even open up new fields of research. For example, recently it was found that mammalian retropseudogenes frequently seem to encode small interfering RNAs, which are important for the regulation of their parental source genes[99,100]. Thus, even retropseudogenes do not necessarily represent evolutionary dead-ends, but might provide the raw material for functionally important evolutionary innovations.

1. Long, M., Betran, E., Thornton, K. & Wang, W. The origin of new genes: glimpses from the young and old. *Nature Rev. Genet.* **4**, 865–875 (2003).
2. Ohno, S. *Evolution by Gene Duplication* (Springer Verlag, Berlin, 1970).
3. Wolfe, K. H. & Li, W. H. Molecular evolution meets the genomics revolution. *Nature Genet.* **33** (Suppl.), 255–265 (2003).
4. Prince, V. E. & Pickett, F. B. Splitting pairs: the diverging fates of duplicated genes. *Nature Rev. Genet.* **3**, 827–837 (2002).
5. Lynch, M. *The Origins of Genome Architecture* (Sinauer Associates, Sunderland, USA 2007).
6. Karin, M. & Richards, R. I. Human metallothionein genes — primary structure of the metallothionein-II gene and a related processed gene. *Nature* **299**, 797–802 (1982).
7. Ueda, S., Nakai, S., Nishida, Y., Hisajima, H. & Honjo, T. Long terminal repeat-like elements flank a human immunoglobulin epsilon pseudogene that lacks introns. *EMBO J.* **1**, 1539–1544 (1982).
8. Hollis, G. F., Hieter, P. A., McBride, O. W., Swan, D. & Leder, P. Processed genes: a dispersed human immunoglobulin gene bearing evidence of RNA-type processing. *Nature* **296**, 321–325 (1982).
9. Petrov, D. A., Lozovskaya, E. R. & Hartl, D. L. High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384**, 346–349 (1996).
10. Jeffs, P. & Ashburner, M. Processed pseudogenes in *Drosophila*. *Proc. Biol. Sci.* **244**, 151–159 (1991).
11. Zhang, Z., Carriero, N. & Gerstein, M. Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet.* **20**, 62–67 (2004).
12. Zhang, Z. L., Harrison, P. M., Liu, Y. & Gerstein, M. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.* **13**, 2541–2558 (2003).
13. Mighell, A. J., Smith, N. R., Robinson, P. A. & Markham, A. F. Vertebrate pseudogenes. *FEBS Lett.* **468**, 109–114 (2000).
14. McCarrey, J. R. & Thomas, K. Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. *Nature* **326**, 501–505 (1987). **The first description of a functional out-of-the-X retrogene.**
15. Bai, Y. S., Casola, C., Feschotte, C. & Betran, E. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in *Drosophila*. *Genome Biol.* **8**, R11 (2007).
16. Betran, E., Thornton, K. & Long, M. Retroposed new genes out of the X in *Drosophila*. *Genome Res.***12**, 1854–1859 (2002).
17. Potrzebowski, L. *et al.* Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol.* **6**, e80 (2008). **The authors generalize the idea that autosomal retrogenes compensate for the silencing of their X-linked parental genes during and after meiosis. Dating of the origin of the out-of-the-X movement pattern of retrogenes revealed that our sex chromosomes are younger than previously thought.**
18. Vinckenbosch, N., Dupanloup, I. & Kaessmann, H. Evolutionary fate of retroposed gene copies in the human genome. *Proc. Natl Acad. Sci. USA* **103**, 3220–3225 (2006). **This study shows that retrocopy transcription is widespread, predominant in the testis and often relies on regulatory elements from nearby genes. It also provides an estimate of the number of functional retrogenes in the human genome.**
19. Emerson, J. J., Kaessmann, H., Betran, E. & Long, M. Y. Extensive gene traffic on the mammalian X chromosome. *Science* **303**, 537–540 (2004). **This global survey of retroposition in human and mouse genomes reveals that the X chromosome has both produced and accepted an excess of retrogenes, thus demonstrating a similar pattern to that previously discovered in fruitflies.**
20. Sayah, D. M., Sokolskaja, E., Berthoux, L. & Luban, J. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature* **430**, 569–573 (2004). **This study shows that retroposition can lead to the fusion of genes with highly complementary functions (in this case, an antiviral function). Strikingly, follow-up studies revealed that this fusion occurred independently in several primate lineages.**
21. Zhang, J., Dean, A. M., Brunet, F. & Long, M. Evolving protein functional diversity in new genes of *Drosophila*. *Proc. Natl Acad. Sci. USA* **101**, 16246–16250 (2004).
22. Zhang, J., Long, M. & Li, L. Translational effects of differential codon usage among intragenic domains of new genes in *Drosophila*. *Biochim. Biophys. Acta* **1728**, 135–142 (2005).
23. Burki, F. & Kaessmann, H. Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nature Genet.* **36**, 1061–1063 (2004). **This study describes one of the few ape-specific new genes for which positively selected amino-acid substitutions could be related to functional change and adaptation.**
24. Dai, H. *et al.* The evolution of courtship behaviors through the origination of a new gene in *Drosophila*. *Proc. Natl Acad. Sci. USA* **105**, 7478–7483 (2008). **The first retrogene for which a behavioural phenotype is described (in this case, behaviour related to courtship).**
25. Shemesh, R., Novik, A., Edelheit, S. & Sorek, R. Genomic fossils as a snapshot of the human transcriptome. *Proc. Natl Acad. Sci. USA* **103**, 1364–1369 (2006).
26. Feng, Q., Moran, J. V., Kazazian, H. H. Jr & Boeke, J. D. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**, 905–916 (1996).
27. Mathias, S. L., Scott, A. F., Kazazian, H. H. Jr, Boeke, J. D. & Gabriel, A. Reverse transcriptase encoded by a human transposable element. *Science* **254**, 1808–1810 (1991).
28. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.* **24**, 363–367 (2000). **The authors demonstrate that the L1 enzymatic machinery can generate processed gene copies, suggesting that the large number of retrocopies in mammals is driven by L1 activity.**
29. Wei, W. *et al.* Human L1 retrotransposition: *cis* preference versus *trans* complementation. *Mol. Cell Biol.* **21**, 1429–1439 (2001).
30. Eickbush, T. H. in *Mobile DNA II* (eds Craig, N. L., Craigie, M., Gellert, M. & Lambowitz, A. M.) 813–835 (American Society of Microbiology, Washington, 2002).
31. Jin, Y. K. & Bennetzen, J. L. Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the Bs1 retroelement of maize. *Plant Cell* **6**, 1177–1186 (1994).
32. Wang, W. *et al.* High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell* **18**, 1791–1802 (2006).
33. Haas, N. B., Grabowski, J. M., Sivitz, A. B. & Burch, J. B. Chicken repeat 1 (CR1) elements, which define an ancient family of vertebrate non-LTR retrotransposons, contain two closely spaced open reading frames. *Gene* **197**, 305–309 (1997).
34. Hillier, L. W. *et al.* Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).
35. Lum, R. & Linial, M. L. Tail-to-head arrangement of a partial chicken glyceraldehyde-3-phosphate dehydrogenase processed pseudogene. *J. Mol. Evol.* **45**, 564–570 (1997).
36. Torrents, D., Suyama, M., Zdobnov, E. & Bork, P. A genome-wide survey of human pseudogenes. *Genome Res.* **13**, 2559–2567 (2003).
37. Marques, A. C., Dupanloup, I., Vinckenbosch, N., Reymond, A. & Kaessmann, H. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol.* **3**, 1970–1979 (2005).
38. Ohshima, K. *et al.* 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 (2003).
39. Warren, W. C. *et al.* Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**, 175–183 (2008).
40. Rohozinski, J., Lamb, D. J. & Bishop, C. E. *UTP14c* is a recently acquired retrogene associated with spermatogenesis and fertility in man. *Biol. Reprod.* **74**, 644–651 (2006).
41. Bradley, J. *et al.* An X-to-autosome retrogene is required for spermatogenesis in mice. *Nature Genet.* **36**, 872–876 (2004). **This study and reference 42 demonstrate that the loss of function of an X-derived retrogene leads to severe defects in male meiotic functions in mice.**
42. Rohozinski, J. & Bishop, C. E. The mouse *juvenile spermatogonial depletion* (*jsd*) phenotype is due to a mutation in the X-derived retrogene, *mUtp14b*. *Proc. Natl Acad. Sci. USA* **101**, 11695–11700 (2004).
43. Dass, B. *et al.* Loss of polyadenylation protein tau CstF-64 causes spermatogenic defects and male infertility. *Proc. Natl Acad. Sci. USA* **104**, 20374–20379 (2007).
44. Ehrmann, I. *et al.* Haploinsufficiency for the germ cell-specific nuclear RNA binding protein hnRNP G-T prevents functional spermatogenesis in the mouse. *Hum. Mol. Genet.* **17**, 2803–2818 (2008).
45. Zhou, Q. *et al.* On the origin of new genes in *Drosophila*. *Genome Res.* **18**, 1446–1455 (2008).
46. Dai, H. Z., Yoshimatsu, T. F. & Long, M. Y. Retrogene movement within- and between-chromosomes in the evolution of *Drosophila* genomes. *Gene* **385**, 96–102 (2006).
47. Harrison, P. M., Milburn, D., Zhang, Z., Bertone, P. & Gerstein, M. Identification of pseudogenes in the *Drosophila melanogaster* genome. *Nucleic Acids Res.* **31**, 1033–1037 (2003).
48. Long, M. & Langley, C. H. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. *Science* **260**, 91–95 (1993).
49. Wang, W., Brunet, F. G., Nevo, E. & Long, M. Origin of sphinx, a young chimeric RNA gene in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **99**, 4448–4453 (2002).
50. Kundu, T. K. & Rao, M. R. CpG islands in chromatin organization and gene expression. *J. Biochem.* **125**, 217–222 (1999).
51. Zaiss, D. M. & Kloetzel, P. M. A second gene encoding the mouse proteasome activator PA28beta subunit is part of a LINE1 element and is driven by a LINE1 promoter. *J. Mol. Biol.* **287**, 829–835 (1999).
52. McCarrey, J. R. Nucleotide sequence of the promoter region of a tissue-specific human retroposon: comparison with its housekeeping progenitor. *Gene* **61**, 291–298 (1987).
53. Shiao, M. S., Liao, B. Y., Long, M. & Yu, H. T. Adaptive evolution of the insulin two-gene system in mouse. *Genetics* **178**, 1683–1691 (2008).
54. Soares, M. B. *et al.* RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon. *Mol. Cell Biol.* **5**, 2090–2103 (1985).
55. Okamura, K. & Nakai, K. Retrotransposition as a source of new promoters. *Mol. Biol. Evol.* **25**, 1231–1238 (2008).
56. Sandelin, A. *et al.* Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Rev. Genet.* **8**, 424–436 (2007).
57. Wood, A. J. *et al.* A screen for retrotransposed imprinted genes reveals an association between X chromosome homology and maternal germ-line methylation. *PloS Genet.* **3**, 192–203 (2007).
58. Parker-Katiraee, L. *et al.* Identification of the imprinted *KLF14* transcription factor undergoing human-specific accelerated evolution. *PLoS Genet.* **3**, e65 (2007).
59. Betran, E. & Long, M. *Dntf-2r*, a young *Drosophila* retroposed gene with specific male expression under positive Darwinian selection. *Genetics* **164**, 977–988 (2003).
60. Yoshioka, H., Geyer, C. B., Hornecker, J. L., Patel, K. T. & McCarrey, J. R. *In vivo* analysis of developmentally and evolutionarily dynamic protein–DNA interactions regulating transcription of the *Pgk2* gene during mammalian spermatogenesis. *Mol. Cell Biol.* **27**, 7871–7885 (2007).
61. Krasnov, A. N. *et al.* A retrocopy of a gene can functionally displace the source gene in evolution. *Nucleic Acids Res.* **33**, 6654–6661 (2005).
62. Bailey, J. A., Church, D. M., Ventura, M., Rocchi, M. & Eichler, E. E. Analysis of segmental duplications and genome assembly in the mouse. *Genome Res.* **14**, 789–801 (2004).
63. Bailey, J. A. & Eichler, E. E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature Rev. Genet.* **7**, 552–564 (2006).
64. Betran, E., Wang, W., Jin, L. & Long, M. Y. Evolution of the *Phosphoglycerate mutase* processed gene in human and chimpanzee revealing the origin of a new primate gene. *Mol. Biol. Evol.* **19**, 654–663 (2002).
65. Rosso, L. *et al.* Birth and rapid subcellular adaptation of a hominoid-specific CDC14 protein. *PLoS Biol.* **6**, e140 (2008). **A combination of evolutionary analyses and molecular- and cell-biology experiments show that the subcellular localization of a protein encoded by an ape-specific retrogene changed during evolution owing to the action of positive selection, thus revealing a novel mechanism for the emergence of new gene function.**

66. Yu, H. J. *et al.* Origination and evolution of a human-specific transmembrane protein gene, *c1orf37-dup*. *Hum. Mol. Genet.* **15**, 1870–1875 (2006).
67. Rosso, L., Marques, A. C., Reichert, A. S. & Kaessmann, H. Mitochondrial targeting adaptation of the hominoid-specific glutamate dehydrogenase driven by positive Darwinian selection. *PLoS Genet.* **4**, e1000150 (2008).
68. Marques, A. C., Vinckenbosh, N., Brawand, D. & Kaessmann, H. Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. *Genome Biol.* **9**, R54 (2008).
69. Smith, E. L. in *The Enzymes* (ed. Boyer, P. D.) 293–367 (Academic, New York, 1975).
70. Mastorodemos, V., Zaganas, I., Spanaki, C., Bessa, M. & Plaitakis, A. Molecular basis of human glutamate dehydrogenase regulation under changing energy demands. *J. Neurosci. Res.* **79**, 65–73 (2005).
71. Byun-McKay, S. A. & Geeta, R. Protein subcellular relocalization: a new perspective on the origin of novel genes. *Trends Ecol. Evol.* **22**, 338–344 (2007).
72. Brennan, G., Kozyrev, Y. & Hu, S. L. TRIMCyp expression in Old World primates *Macaca nemestrina* and *Macaca fascicularis*. *Proc. Natl Acad. Sci. USA* **105**, 3569–3574 (2008).
73. Virgen, C. A., Kratovac, Z., Bieniasz, P. D. & Hatziioannou, T. Independent genesis of chimeric TRIM5-cyclophilin proteins in two primate species. *Proc. Natl Acad. Sci. USA* **105**, 3563–3568 (2008).
74. Wilson, S. J. *et al.* Independent evolution of an antiviral TRIMCyp in rhesus macaques. *Proc. Natl Acad. Sci. USA* **105**, 3557–3562 (2008).
75. Akiva, P. *et al.* Transcription-mediated gene fusion in the human genome. *Genome Res.* **16**, 30–36 (2006).
76. Babushok, D. V. *et al.* A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids. *Genome Res.* **17**, 1129–1138 (2007).
77. Parra, G. *et al.* Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res.* **16**, 37–44 (2006).
78. Kleene, K. C. A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. *Mech. Dev.* **106**, 3–23 (2001).
79. She, X. *et al.* The structure and evolution of centromeric transition regions within the human genome. *Nature* **430**, 857–864 (2004).
80. Fontanillas, P., Hartl, D. L. & Reuter, M. Genome organization and gene expression shape the transposable element distribution in the *Drosophila melanogaster* euchromatin. *PLoS Genet.* **3**, e210 (2007).
81. Bai, Y., Casola, C. & Betran, E. Evolutionary origin of regulatory regions of retrogenes in *Drosophila*. *BMC Genomics* **9**, 241 (2008).
82. Wang, P. J. X chromosomes, retrogenes and their role in male reproduction. *Trends Endocrinol. Metab.* **15**, 79–83 (2004).
83. Shiao, M. S. *et al.* Origins of new male germ-line functions from X-derived autosomal retrogenes in the mouse. *Mol. Biol. Evol.* **24**, 2242–2253 (2007).
84. Sturgill, D., Zhang, Y., Parisi, M. & Oliver, B. Demasculinization of X chromosomes in the *Drosophila* genus. *Nature* **450**, 238–241 (2007).
85. Hense, W., Baines, J. F. & Parsch, J. X chromosome inactivation during *Drosophila* spermatogenesis. *PLoS Biol.* **5**, e273 (2007).
86. Lahn, B. T. & Page, D. C. Four evolutionary strata on the human X chromosome. *Science* **286**, 964–967 (1999).
87. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
88. McLysaght, A. Evolutionary steps of sex chromosomes are reflected in retrogenes. *Trends Genet.* **10**, 478–481 (2008).
89. Veyrunes, F. *et al.* Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. *Genome Res.* **18**, 965–973 (2008).
90. Rens, W. *et al.* The multiple sex chromosomes of platypus and echidna are not completely identical and several share homology with the avian Z. *Genome Biol.* **8**, R243 (2007).
91. Ross, M. T. *et al.* The DNA sequence of the human X chromosome. *Nature* **434**, 325–337 (2005).
92. Betran, E., Emerson, J. J., Kaessmann, H. & Long, M. Sex chromosomes and male functions — where do new genes go? *Cell Cycle* **3**, 873–875 (2004).
93. *Roy.*, S. W. & Gilbert, W. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nature Rev. Genet.* **7**, 211–221 (2006).
94. Coulombe-Huntington, J. & Majewski, J. Characterization of intron loss events in mammals. *Genome Res.* **17**, 23–32 (2007).
95. Fink, G. R. Pseudogenes in yeast? *Cell* **49**, 5–6 (1987).
96. Goffeau, A. *et al.* Life with 6,000 genes. *Science* **274**, **546**, 563–567 (1996).
97. Parmley, J. L., Urrutia, A. O., Potrzebowski, L., Kaessmann, H. & Hurst, L. D. Splicing and the evolution of proteins in mammals. *PloS Biol.* **5**, 343–353 (2007).
98. Kaiser, J. DNA sequencing. A plan to capture human diversity in 1,000 genomes. *Science* **319**, 395 (2008).
99. Tam, O. H. *et al.* Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**, 534–538 (2008).
100. Watanabe, T. *et al.* Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453**, 539–543 (2008).
101. Pain, D., Chirn, G. W., Strassel, C. & Kemp, D. M. Multiple retropseudogenes from pluripotent cell-specific gene expression indicates a potential signature for novel gene identification. *J. Biol. Chem.* **280**, 6265–6268 (2005).
102. Huang, Y. T., Chen, F. C., Chen, C. J., Chen, H. L. & Chuang, T. J. Identification and analysis of ancestral hominoid transcriptome inferred from cross-species transcript and processed pseudogene comparisons. *Genome Res.* **18**, 1163–1170 (2008).
103. Wang, P. J. & Page, D. C. Functional substitution for TAF(II)250 by a retroposed homolog that is expressed in human spermatogenesis. *Hum. Mol. Genet.* **11**, 2341–2346 (2002).
104. Nielsen, R. *et al.* A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3**, e170 (2005).
105. Jones, C. D. & Begun, D. J. Parallel evolution of chimeric fusion genes. *Proc. Natl Acad. Sci. USA* **102**, 11373–11378 (2005).
106. Kalamegham, R., Sturgill, D., Siegfried, E. & Oliver, B. *Drosophila* mojoless, a retroposed GSK-3, has functionally diverged to acquire an essential role in male fertility. *Mol. Biol. Evol.* **24**, 732–742 (2007).

Acknowledgements

We apologize to colleagues whose work could not be discussed or cited owing to space constraints and/or the focus of this Review. We thank the members of the H.K. and M.L. laboratories for discussions. This work was supported by funds from the Swiss National Science Foundation, the European Research Council (STREP: 140404), and EMBO Young Investigator Grant (to H.K.), as well as the National Institutes of Health (R0IGM078070-01A1) (to M. L.).

DATABASES

Entrez Gene:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
CDC14B | *GLUD1* | *GLUD2* | *PIPSL* | *TRIM5*

FURTHER INFORMATION

Henrik Kaessmann's homepage:
http://www.unil.ch/cig/page7858_en.html
Manyuan Long's homepage:
http://pondside.uchicago.edu/ecol-evol/faculty/long_m.html

ALL LINKS ARE ACTIVE IN THE ONLINE PDF