Title: RNA-based gene duplication: mechanistic and evolutionary insights.

Authors: Kaessmann H, Vinckenbosch N, Long M

Journal: Nature reviews. Genetics

Year: 2009 Jan

Volume: 10

Issue: 1

Pages: 19-31

DOI: 10.1038/nrg2487

Published in final edited form as:

Nat Rev Genet. 2009 January ; 10(1): 19–31. doi:10.1038/nrg2487.

RNA-based gene duplication: mechanistic and evolutionary insights

Henrik Kaessmann[1], **Nicolas Vinckenbosch**[1], and **Manyuan Long**[2]

[1]Center for Integrative Genomics, University of Lausanne, Genopode, CH-1015 Lausanne, Switzerland. [2]Department of Ecology and Evolution, The University of Chicago, 1101 East 57th Street, Chicago, Illinois 60637, USA.

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 but also some non-mammalian genomes (in particular that of the fruitfly). These studies not only revealed previously unknown mechanisms for the emergence of new genes and their functions but 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 new light on the evolutionary origin and biology of our sex chromosomes.

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 (i.e. duplication of chromosomal segments containing genes) have been considered and widely studied in this context (reviewed e.g. in refs [4,5]), although 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" (ref. [9-12]) and routinely classified as processed pseudogenes[13] due to the expected lack of regulatory elements and presence of mutations in many copies such as premature stop codons. Indeed, they were mainly considered a nuisance and confounding factor in transcription surveys because of their often high sequence similarity with parental source genes.

However, after some anecdotal findings of functional retroposed genes since the late 1980s (e.g. ref. [14]), an unexpectedly large number of functional retrogenes have recently been discovered - mainly in mammals and fruitflies (e.g. refs [15-19]). These studies revealed that retrogenes often evolved functional roles in the male germline (e.g. ref. [16,17]), while other intriguing retrogene functions - e.g. in anti-viral defense[20], in hormone-pheromone metabolism[21,22], in the brain[23], or in courtship behaviors[24] – have also been postulated. More fundamentally, retrogene analyses have uncovered novel mechanisms with respect to how new genes may arise (e.g. the recruitment of regulatory elements) and obtain new functions (e.g. through gene fusion and adaptive evolution). Finally, retroposed gene copies

Correspondence to H.K. (Henrik.Kaessmann@unil.ch).

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 only possible thanks to the growing number of complete genome sequences and achieved by targeted cross-disciplinary approaches, which involved evolutionary analysis, mining of available large-scale expression data, and molecular/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 retroduplication is most abundant and/or best studied in mammals and fruitflies, we will focus our discussion on these organisms. Specifically, after briefly introducing the process of retroduplication, we first discuss the abundance of retrocopies and functional retrogenes in mammals and *Drosophila*. We then proceed with a discussion on how retrocopies may become transcribed and functional, which is followed by an overview of novel mechanisms underlying the emergence of new gene functions that were uncovered in detailed surveys of young retrogenes. We then discuss a major functional role of retrogenes in the male germline, which is related to the biology and evolution of X chromosomes. Finally, we round off the review with a discussion of other general insights pertaining to mammalian genome evolution obtained from global retrocopy surveys and some concluding remarks on potential future research directions.

Mechanisms of retroposition

To be heritable and hence of evolutionary relevance, retroposition/retroduplication (these terms are used interchangeably) needs to occur in the germline (or during early embryonic stages). Thus, retroposition requires an 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 germline. The fact that retroposition relies on the duplication through an mRNA intermediate also implies that only genes expressed in the germline can be duplicated via this mechanism.

The key retroduplication enzyme, reverse transcriptase, appears to generally 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 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 Figure 1.

Retrotransposable element-encoded enzymes are likely also responsible for retroposition in *Drosophila*[10,30] and some plants[31,32], which carry various retrotransposons with reverse transcriptase activity, although the retroposition machinery has not been studied in detail in these organisms to date. The paucity of retrocopies in non-mammalian vertebrates is likely 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 LINE elements[33], but CR1 reverse transcriptases cannot recognize polyadenylated mRNAs (due to their specificity towards a different target sequence) and are thus incapable of promoting retroposition of mRNAs from genes in the genome[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] (due to the high activity of L1 LINE elements), we first discuss the rates of retrocopy and functional retrogene formation in mammals and then in *Drosophila*. Thousands of retrocopies have been identified in several placental mammal (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], likely due to the fluctuating activity of L1 elements (see also BOX 1). Recently, approximately 2000 retrocopies were identified in the opossum genome[17], suggesting a similarly high retroposition rate in metatherians (marsupials). Only few retrocopies (in the order of 50) seem to be present in the platypus genome (Soumillon et al., unpublished), consistent with the paucity of L1 elements in monotremes[39], the most basal mammalian lineage.

It was long assumed that retroposed gene copies represent, by and large, non-functional retropseudogenes due to their presumed lack of expression potential[10,13], although individual studies have revealed instances of functional retrogenes since the late eighties[14]. But how many retrocopies have evolved into *bona fide* genes? Different types of evidence can be used to support functionality of retrocopies. Given the wealth of genomic data, rather straightforward approaches to support retrogene functionality are based on evolutionary analyses that screen for signatures of selection. For example, the (selective) preservation of intact open reading frames (ORFs) between distant[17,18] or several closely related species[37] provides statistically significant and convincing evidence for non-neutral evolution of retrocopies and therefore their functionality. Furthermore, comparisons of the rates of functionally relevant (amino acid changing) substitutions and neutral changes (silent substitutions) in retrogene coding regions can be used to detect non-neutral evolution, indicative of functional constraint (e.g. refs [23,37]). In addition to such evolutionary approaches, molecular signs of functionality may be sought, such as evidence for transcription, which can often be readily obtained. But on its own, this does not suffice to support functionality of individual genes, as non-functional DNA (including retropseudogenes[18]) might be transcribed as well. Evidence for translation, that is, the presence of a protein (e.g. 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 have usually been obtained based on evolutionary/statistical analyses.

Vinckenbosch et al. estimated the number of functional retrogenes present in the human genome by comparing transcription levels of intact retrocopies to those of retropseudogenes, which reflect the transcriptional background noise in the genome[18]. The authors showed 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 then conservatively estimated that at least 120 retrocopies are likely to represent 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 leading to humans[37].

In *Drosophila*, where 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 invertebrate lineage are functional[15,16,46]. However, the total number of retrocopies in this genus is much lower than that in mammals, which seems mainly to be due to the paucity of retropseudogenes in the *Drosophila* genome,[9,47] (because of the extremely short half-life of unconstrained DNA in this genus[9]), rather than 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 how retrocopies can be expressed in their new genomic location. To become expressed at a significant level and in a meaningful way (e.g. in tissues where it can exert a selectively beneficial function), a new gene needs to obtain a core promoter and probably other elements (e.g. enhancers) that regulate its expression. In the following, we discuss various mechanisms through which the acquisition of promoters and other regulatory elements may occur.

Generally, retrocopies may profit from pre-exisiting regulatory elements in their vicinity for their expression. For example, a straightforward way for a retrocopy to obtain transcription potential would be to directly hitchhike on the regulatory machinery of other genes. Indeed, a number of cases have been described where retrocopies are located in an intron of a host gene, being 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 only with 5′ untranslated (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 may be facilitated by the open chromatin and/or regulatory elements of nearby genes[18] (Fig. 2B). The latter possibility is supported by observations that retrogenes may be transcribed from bi-directional CpG-rich promoters of genes in their proximity (Fablet et al., manuscript in preparation). 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 (i.e. via the acquisition of new 5′ UTR structures), retrocopies may also become transcribed from distant CpG-enriched sequences (which often have inherent capacity to promote transcription[50]) not previously associated with other genes (Fig. 2C; Fablet et al., manuscript in preparation). These distant CpG “proto-promoter” elements may have been optimized by natural selection once associated with a functional retrogene. Similarly, distant promoters from retrotransposable elements have been “captured” by retrocopies for their transcription via the acquisition of new 5′ untranslated exon/intron structures (Fablet et al., manuscript in preparation). In addition, retrotransposons[51] (or potentially CpG island proto-promoters) immediately upstream of retrogene insertion sites may also be used directly (Fig. 2D).

Until recently, it was thought that retrocopies are quite 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 might nevertheless rather 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] (TSS). If a retrocopy stems from a parental transcript with a TSS located relatively far upstream, the mRNA that gave rise to the retrocopy may 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*, the source of transcription potential of retrogenes is somewhat more elusive. While – similarly to mammals - host gene fusions have occurred in this genus (e.g. refs [48,49]) and retrogene transcription may 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 role in the formation of basic *Drosophila* retrogene promoters[15,59] (Fig. 2F).

We note that the various mechanisms that may endow retrogenes with regulatory elements described here probably often only provide the basic means for the initial transcription of retrocopies, while more sophisticated regulatory elements may evolve with time (see e.g. the mammalian *Pgk2* retrogene; Table 1; ref. [52,60]).

The evolution of new functions from retrogenes

DNA versus RNA-based duplication

The fundamental differences between the two major duplication mechanisms – segmental duplication (reviewed e.g. in refs [4,5]) and retroposition – have significant consequences for the respective evolutionary fates of generated 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 retroduplicate copies usually lack introns and are less likely to have strong regulatory elements upon their emergence. Therefore, segmental duplication is more likely to yield expressed daughter copies than the retroduplication process. At the same time, segmental duplicates are likely to exhibit very similar expression patterns in their early evolution, which may often imply that one copy is initially functionally redundant, and the increased gene dose might even deleterious (although increased gene dosage may sometimes be beneficial and thus selectively preserved). By contrast, retroduplicate copies often need to recruit regulatory elements to become transcribed (see section above). This also means, however, 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 arising 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 retroduplication process (often not discernible for segmental duplications) facilitates studies pertaining to the origin of new gene functions, since parental genes usually maintain the ancestral gene function (although there are interesting exceptions to this rule, ref. [61]), while new functions usually are acquired by the intronless daughter retrogene copies. It also renders the detection and analysis of young duplication events, which are particularly informative for the study of new gene functions (see below), straightforward. Recent segmental duplicates, on the other hand, are not easily distinguishable and more difficult to study, as they are, for example, frequently collapsed into a single locus in standard genome assemblies due to their high sequence and structural similarities.

Finally, retroduplication usually produces gene copies on chromosomes different from that of the parental gene copy, while segmental duplications are less likely to involve different chromosomes (although the rate of inter- vs. intrachromsomal segmental duplication differs between lineages, refs [45,62,63]). Thus, retroduplication represents 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, due to the abundance of functional segmental duplicates in nearly all studied genomes, numerous studies of segmental duplication have yielded many fundamental insights and established general concepts regarding the emergence of new gene functions (reviewed in detail in e.g. refs [4,5]),

However, due to the particular features of retroposed gene copies outlined above, the analysis of retroduplication 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 recently 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).

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 collectively termed subcellular adaptation[65,67,68]. The following two examples led to the finding of subcellular adaptation and demonstrate two ways by which this process might occur (Fig. 3).

The study of the *GLUD2* retrogene exemplifies one form of subcellular adaptation (“sublocalization”, ref. [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 MYA by retroposition from its parental gene *GLUD1*, which encodes an enzyme that degrades glutamate[69]. The *GLUD2*-encoded enzyme evolved unique biochemical properties soon after the duplication event by virtue of 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 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 to one of these compartments, the mitochondrion, due to a single, positively selected substitution in its N-terminal targeting sequence[67]. This event likely contributed to the adaptation of GLUD2 to a function in the glutamate metabolism of the brain and other tissues. Thus, GLUD2 represents 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. 3).

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, reflecting a variant form of subcellular adaptation that was termed subcellular relocalization or neolocalization[68,71]. *CDC14Bretro* stems from a splice variant of the *CDC14B* cell cycle gene[65] (Table 1) and encodes a protein that became specifically expressed in the adult/fetal brain and testes soon after its emergence in the common human and ape ancestor. It then completely relocalized in the cell due to intense positive selection in the common African ape ancestor ~7-12 Mya, shifting from the ancestral association with microtubules (which it stabilized) to a localization and function on the endoplasmic reticulum (Fig. 3).

Notably, a recent global survey of yeast duplicate proteins, prompted by these retrogene studies, showed that subcellular adaptation appears to be widespread, being 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[5] (through neo- or subfunctionalization), rapid and selectively driven subcellular adaptation by either “neolocalization” (CDC14Bretro) or “sublocalization” (GLUD2) represents a common, previously little considered mechanism underlying the emergence of new gene function (Fig. 3).

Gene fusion and domain shuffling

Another way by which new gene functions can arise is through gene fusion, which is defined as the fusion of two previously separate source genes into a single transcription unit[1]. Gene fusion may 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 represents a form of exon or domain shuffling[1].

Fusions of retroposed gene copies to genes into which they insert have yielded new genes with important functions. Detailed studies of such fusion genes uncovered surprising aspects of new gene formation such as the recurrent juxtaposition of genes with complementary functions, as in the case of the *TRIM5-CypA* fusion gene (Fig. 4). A retroposed copy of the *CypA* gene, whose encoded protein potently binds retroviral capsids, was shown to have integrated independently into the antiviral defense gene *TRIM5* in a New World monkey[20] (Fig. 4A) and 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 newly emerged TRIM5-CypA fusion protein more efficiently restricts HIV-1 and other retroviruses in these species[20,72-74]. The *TRIM5-CypA* gene fusion represents a striking case of domain shuffling and convergent evolution. The at first glance seemingly unlikely multiple independent insertions of *CypA* retrocopies into the same gene were probably facilitated by a rather high retroposition rate of the *CypA* gene (due to its high expression in the germline). Rare *TRIM5-CypA* fusions were then likely 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 recent retroposed gene (*PIPSL*) on human chromosome 10 that stems from a fusion transcript of two parental genes (*PIP5K1A* and *PSMD4*) that reside adjacently 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 appeared to be post-transcriptionally repressed. However, in a recent follow-up analysis, we (manuscript submitted) obtained evolutionary and experimental support for the functionality of this gene in hominoids. 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* 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 represents the first chimeric gene involving retroposition described in any species[48], originated by the insertion of a retrocopy of the Alcohol dehydrogenase gene (*Adh*) into the *yande* gene[48] (Table 1). The functional evolution of jgw was recently unveiled using a biochemical approach2122, which revealed that the JGW protein was shaped by positive selection (in particular the ADH domain) and apparently evolved a role in hormone/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 as yet poorly understood. *Sphinx* 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 (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]. The phenotype of these *spx* knockout flies – increased male-male courtship behaviour relative to wild type *Drosophila* – suggests that *spx* represents the first recently emerged gene for which a behavioral phenotype could be identified.

Retrogene functions in testes and sex chromosome evolution

In the following, we will discuss global surveys of retroposition in mammals and fruitflies, which have shown that retrogenes often evolved functions in the testes and that the formation and preservation of many of these genes is closely linked to the biology and selective forces (imposed by the male germline) that have shaped X chromosomes ever since their emergence. Dating of the origin of these retrogenes also allowed a reassessment of the age of mammalian sex chromosomes.

Expression in testes

Numerous retrogene studies in both mammals and fruitflies revealed an overall propensity of retrogenes to be expressed in testes (refs [16,18,37,46,48] and references therein). 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 appear to be in a state of hypertranscription due to various modifications of the chromatin (reviewed in ref. [78]). This hypertranscription state was suggested to allow transcription of DNA that is usually not transcribed and therefore might have facilitated transcription of retrocopies[37] but also of other types of duplicates[79] in testis during their early evolution. A subset of these retrocopies subsequently obtained beneficial functions in testis and evolved into *bona fide* genes (see further discussion 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 retroduplication occurs in the germline, they might therefore predominantly insert into or close to germline-expressed genes, which would facilitate retrocopy transcription in the germline. However, in *Drosophila*, this hypothesis appears to explain testis expression of only some retrogenes (several retrogenes are located in regions with many testis-expressed genes, ref. [81]). In mammals, this insertion bias scenario remains to be explored.

Retrogenes out of the X

As pointed out above, the retroduplication 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 *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 retrogene are specifically expressed in 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 (Fig. 5A), due to male meiotic sex chromosome inactivation (MSCI) (ref. [17] and studies reviewed in ref. [82]).

Importantly, these mammalian X-derived retrogenes are significantly more frequently and more specifically expressed during and post meiosis than other retrogenes[17] (which also tend to be expressed in testes – see subsection above). This substantiates the hypothesis that retrogenes that stem from the X 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 functionally supported by 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 humans[40]. It is worth pointing out that, curiously, the potential mechanistic biases favoring expression in meiotic/post-meiotic cells (see subsection above) allow X-derived retrogenes to be expressed precisely where needed to compensate their parents. Thus, together with the fact that the retroduplication process readily moves genes between chromosomes, this means that retrogenes – rather than DNA-based duplicates – may 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-X movement in *Drosophila* might be different from that in mammals[84], a recent study suggests that MSCI may occur in *Drosophila* (ref. [85]). Therefore, MSCI may be the main force responsible for the preferential fixation of X-derived retrogenes with meiotic/post-meiotic expression in fruitflies as well. 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 primate retrogenes showed that the out-of-X movement of retrogenes is ongoing[37], which suggests that gene export from the X 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 were generated ever since the common ancestor of all mammals, selectively driven retrogene export from the X only started later, on the eutherian and marsupial lineages, respectively[17] (Fig. 5B). Given that MSCI is the likely selective force driving genes off the X, this observation suggested that MSCI emerged – rather late - in the common ancestor of eutherians and marsupials, that is, well after their separation from the monotreme lineage[17] (Fig. 5B).

Moreover, these observations lead 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 late) 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 (eutherian/marsupial) sex chromosomes[39,89,90].

Retroposition into the X

Curiously, retrogenes are not only exported from the X but are also prefentially imported into this chromosome in mammals (ref. [19]). There seems to exist a slight mechanistic bias that favors the insertion and/or retention of retrocopies on the X (ref. [19]). Although the cause of this bias remains unclear, the excess of retropseudogenes on the X in is consistent with the accumulation of other non-functional retro-elements (LINEs) on the X chromosome in this lineage[91]. In addition, however, a strong selective force - the precise nature of which remains to be identified - apparently led to the preferential fixation of *bona fide* retrogenes on the X (ref. [19]). We finally note that no increased fixation rate of retrogenes on the X is observed in *Drosophila*[16,92]. This may reflect differences in the biology of sex chromosomes between mammals and fruitflies, but the precise reasons for this discrepancy needs to be clarified.

Retrocopies 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. We discuss these findings in the following subsections and in BOX 1, which highlights how retrocopies reflect aspects of transcriptome evolution.

Retrocopies and intron loss

One way by which retrocopies have shaped mammalian genes is by mediating the loss of introns. Intron gains are rare events during evolution, while intron loss appears 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]. Interestingly, these intron losses appear to have been mediated by recombination of the gene displaying intron loss with the reverse-transcribed, processed mRNA molecule (cDNA) of the same gene[94,95]. There are several lines of evidence supporting this hypothesis, including 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 germline (thus producing many processed cDNAs that may 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 (enhancer/silencers) near exon boundaries significantly reduces the rate of protein evolution[97]. The rate of protein evolution of retrogenes is highest near the sequences where intron-exon junctions previously resided in the parental genes that gave rise to the retrogenes. Therefore, splicing sequence constraints may 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 to their intron-containing parents due the relaxation of splicing constraints.

Conclusions

Messenger RNA-derived duplicates were long thought to be doomed to pseudogenization and decay. As outlined in this review, however, a significant number of retroposed gene copies have escaped this evolutionary fate and have evolved into *bona fide* genes. Retroduplicate genes are probably still much less likely to become functional compared to “normal” DNA duplicates due to their peculiar properties, which include the frequent lack of strong regulatory elements upon their emergence. On the other hand, due to these properties, retrogenes often evolved in unique ways, being much more prone to evolve new expression patterns, new genomic locations, and new functions than DNA duplicates. Thus, individual and global surveys of retrogenes (using a variety of evolutionary, genomics, and molecular tools) have unearthed previously unknown molecular mechanisms pertaining to the origin of new genes (e.g. promoter recruitment, subcellular adaptation of encoded proteins), and have provided unexpected and unique insights into genome evolution (e.g. 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. To date, only relatively few young retrogenes have been pinpointed and even fewer studies (most of them 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), due to the difficulty in assessing their selective preservation, but will benefit from the steadily increasing number of available complete genomes in primates. Notably, very recent functional hominoid retrocopies might soon be identified based on an astounding number of human genomes that will soon be completed using the new, 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 combinations of evolutionary analysis with molecular, cellular, and *in vivo* experiments (e.g. 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 genomes from 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 them fused to other genes), an unexpected finding, given that the retroposition activity in plants was traditionally thought to be low.

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

Acknowledgments

We apologize to colleagues whose work could not be discussed or cited due to space constraints and/or the focus of this review. We thank the members of the H.K. and M.L. laboratories for helpful 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 NIH (R0IGM078070-01A1) (to M. L.).

Glossary

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, retroduplication).

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

RETROCOPY Gene copy that results from the process of retroposition (also termed retroposed gene copy, retroduplicate copy).

RETROGENE Expressed and functional retrocopy (usually with an intact open reading frame consistent with that of the parental gene).

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

L1 ELEMENTS A member of the long interspersed retrotransposable (LINE) family of repeats, which provides the enzymatic machinery necessary for the process of retroposition.

NEW GENE A gene that originated recently during evolution.

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.

GENE FUSION The fusion of adjacent genes into a single transcription unit (termed chimeric gene or fusion gene).

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

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

References

1. Long M, Betran E, Thornton K, Wang W. The origin of new genes: Glimpses from the young and old. Nature Reviews Genetics. 2003; 4:865–875.
2. Ohno, S. Evolution by Gene Duplication. Springer Verlag; Berlin: 1970.
3. Wolfe KH, Li WH. Molecular evolution meets the genomics revolution. Nat Genet. 2003; 33(Suppl):255–65. [PubMed: 12610535]
4. Prince VE, Pickett FB. Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet. 2002; 3:827–37. [PubMed: 12415313]
5. Lynch, M. The origins of genome architecture. Sinauer Associates; Sunderland, USA: 2007.
6. Karin M, Richards RI. Human metallothionein genes--primary structure of the metallothionein-II gene and a related processed gene. Nature. 1982; 299:797–802. [PubMed: 7133118]
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. 1982; 1:1539–44. [PubMed: 6327276]

8. Hollis GF, Hieter PA, McBride OW, Swan D, Leder P. Processed genes: a dispersed human immunoglobulin gene bearing evidence of RNA-type processing. Nature. 1982; 296:321–5. [PubMed: 6801526]

9. Petrov DA, Lozovskaya ER, Hartl DL. High intrinsic rate of DNA loss in Drosophila. Nature. 1996; 384:346–9. [PubMed: 8934517]

10. Jeffs P, Ashburner M. Processed pseudogenes in Drosophila. Proc Biol Sci. 1991; 244:151–9. [PubMed: 1679549]

11. Zhang Z, Carriero N, Gerstein M. Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet. 2004; 20:62–7. [PubMed: 14746985]

12. Zhang ZL, Harrison PM, Liu Y, Gerstein M. Millions of years of evolution preserved: A comprehensive catalog of the processed pseudogenes in the human genome. Genome Research. 2003; 13:2541–2558. [PubMed: 14656962]

13. Mighell AJ, Smith NR, Robinson PA, Markham AF. Vertebrate pseudogenes. FEBS Lett. 2000; 468:109–14. [PubMed: 10692568]

14. McCarrey JR, Thomas K. Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. Nature. 1987; 326:501–505. [PubMed: 3453121]

15. Bai YS, Casola C, Feschotte C, Betran E. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila. Genome Biology. 2007; 8

16. Betran E, Thornton K, Long M. Retroposed new genes out of the X in Drosophila. Genome Research. 2002; 12:1854–1859. [PubMed: 12466289]

17. Potrzebowski L, et al. Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS Biol. 2008; 6:e80. [PubMed: 18384235]

18. Vinckenbosch N, Dupanloup I, Kaessmann H. Evolutionary fate of retroposed gene copies in the human genome. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103:3220–3225. [PubMed: 16492757]

19. Emerson JJ, Kaessmann H, Betran E, Long MY. Extensive gene traffic on the mammalian X chromosome. Science. 2004; 303:537–540. [PubMed: 14739461]

20. Sayah DM, Sokolskaja E, Berthoux L, Luban J. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. Nature. 2004; 430:569–73. [PubMed: 15243629]

21. Zhang J, Dean AM, Brunet F, Long M. Evolving protein functional diversity in new genes of Drosophila. Proc Natl Acad Sci U S A. 2004; 101:16246–50. [PubMed: 15534206]

22. Zhang J, Long M, Li L. Translational effects of differential codon usage among intragenic domains of new genes in Drosophila. Biochim Biophys Acta. 2005; 1728:135–42. [PubMed: 15833448]

23. Burki F, Kaessmann H. Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. Nature Genetics. 2004; 36:1061–1063. [PubMed: 15378063]

24. Dai H, et al. The evolution of courtship behaviors through the origination of a new gene in Drosophila. Proc Natl Acad Sci U S A. 2008; 105:7478–83. [PubMed: 18508971]

25. Shemesh R, Novik A, Edelheit S, Sorek R. Genomic fossils as a snapshot of the human transcriptome. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103:1364–1369. [PubMed: 16432206]

26. Feng Q, Moran JV, Kazazian HH Jr, Boeke JD. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell. 1996; 87:905–16. [PubMed: 8945517]

27. Mathias SL, Scott AF, Kazazian HH Jr. Boeke JD, Gabriel A. Reverse transcriptase encoded by a human transposable element. Science. 1991; 254:1808–10. [PubMed: 1722352]

28. Esnault C, Maestre J, Heidmann T. Human LINE retrotransposons generate processed pseudogenes. Nat Genet. 2000; 24:363–7. [PubMed: 10742098]

29. Wei W, et al. Human L1 retrotransposition: cis preference versus trans complementation. Mol Cell Biol. 2001; 21:1429–39. [PubMed: 11158327]

30. Eickbush, TH. Mobile DNA II. Craig, NL.; Craigie, M.; Gellert, M.; Lambowitz, AM., editors. American Society of Microbiology Press; Washington, DC: 2002. p. 813-835.

31. Jin YK, Bennetzen JL. Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the Bs1 retroelement of maize. Plant Cell. 1994; 6:1177–86. [PubMed: 7919987]

32. Wang W, et al. High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell. 2006; 18:1791–802. [PubMed: 16829590]

33. Haas NB, Grabowski JM, Sivitz AB, Burch JB. Chicken repeat 1 (CR1) elements, which define an ancient family of vertebrate non-LTR retrotransposons, contain two closely spaced open reading frames. Gene. 1997; 197:305–9. [PubMed: 9332379]

34. Hillier LW, et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004; 432:695–716. [PubMed: 15592404]

35. Lum R, Linial ML. Tail-to-head arrangement of a partial chicken glyceraldehyde-3-phosphate dehydrogenase processed pseudogene. J Mol Evol. 1997; 45:564–70. [PubMed: 9342403]

36. Torrents D, Suyama M, Zdobnov E, Bork P. A genome-wide survey of human pseudogenes. Genome Research. 2003; 13:2559–2567. [PubMed: 14656963]

37. Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. Emergence of young human genes after a burst of retroposition in primates. Plos Biology. 2005; 3:1970–1979.

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 Biology. 2003; 4

39. Warren WC, et al. Genome analysis of the platypus reveals unique signatures of evolution. Nature. 2008; 453:175–83. [PubMed: 18464734]

40. Rohozinski J, Lamb DJ, Bishop CE. UTP14c is a recently acquired retrogene associated with spermatogenesis and fertility in man. Biology of Reproduction. 2006; 74:644–651. [PubMed: 16354793]

41. Bradley J, et al. An X-to-autosome retrogene is required for spermatogenesis in mice. Nature Genetics. 2004; 36:872–876. [PubMed: 15258580]

42. Rohozinski J, Bishop CE. The mouse juvenile spermatogonial depletion (jsd) phenotype is due to a mutation in the X-derived retrogene, mUtp14b. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101:11695–11700. [PubMed: 15289605]

43. Dass B, et al. Loss of polyadenylation protein tau CstF-64 causes spermatogenic defects and male infertility. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104:20374–20379. [PubMed: 18077340]

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. 2008; 17:2803–2818. [PubMed: 18562473]

45. Zhou Q, et al. On the origin of new genes in Drosophila. Genome Res. 2008; 18:1446–1455. [PubMed: 18550802]

46. Dai HZ, Yoshimatsu TF, Long MY. Retrogene movement within- and between-chromosomes in the evolution of Drosophila genomes. Gene. 2006; 385:96–102. [PubMed: 17101240]

47. Harrison PM, Milburn D, Zhang Z, Bertone P, Gerstein M. Identification of pseudogenes in the Drosophila melanogaster genome. Nucleic Acids Research. 2003; 31:1033–1037. [PubMed: 12560500]

48. Long M, Langley CH. Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science. 1993; 260:91–5. [PubMed: 7682012]

49. Wang W, Brunet FG, Nevo E, Long M. Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. Proc Natl Acad Sci U S A. 2002; 99:4448–53. [PubMed: 11904380]

50. Kundu TK, Rao MR. CpG islands in chromatin organization and gene expression. J Biochem. 1999; 125:217–22. [PubMed: 9990116]

51. Zaiss DM, Kloetzel PM. 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. 1999; 287:829–35. [PubMed: 10222192]

52. McCarrey JR. Nucleotide sequence of the promoter region of a tissue-specific human retroposon: comparison with its housekeeping progenitor. Gene. 1987; 61:291–8. [PubMed: 3446575]

53. Shiao MS, Liao BY, Long M, Yu HT. Adaptive evolution of the insulin two-gene system in mouse. Genetics. 2008; 178:1683–91. [PubMed: 18245324]

54. Soares MB, et al. RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon. Mol Cell Biol. 1985; 5:2090–103. [PubMed: 2427930]

55. Okamura K, Nakai K. Retrotransposition as a source of new promoters. Mol Biol Evol. 2008; 25:1231–8. [PubMed: 18367464]

56. Sandelin A, et al. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat Rev Genet. 2007; 8:424–36. [PubMed: 17486122]

57. Wood AJ, et al. A screen for retrotransposed imprinted genes reveals an association between X chromosome homology and maternal germ-line methylation. Plos Genetics. 2007; 3:192–203.

58. Parker-Katiraee L, et al. Identification of the imprinted KLF14 transcription factor undergoing human-specific accelerated evolution. PLoS Genet. 2007; 3:e65. [PubMed: 17480121]

59. Betran E, Long M. Dntf-2r, a young Drosophila retroposed gene with specific male expression under positive Darwinian selection. Genetics. 2003; 164:977–88. [PubMed: 12871908]

60. Yoshioka H, Geyer CB, Hornecker JL, Patel KT, McCarrey JR. In vivo analysis of developmentally and evolutionarily dynamic protein-DNA interactions regulating transcription of the Pgk2 gene during mammalian spermatogenesis. Mol Cell Biol. 2007; 27:7871–85. [PubMed: 17875925]

61. Krasnov AN, et al. A retrocopy of a gene can functionally displace the source gene in evolution. Nucleic Acids Res. 2005; 33:6654–61. [PubMed: 16314324]

62. Bailey JA, Church DM, Ventura M, Rocchi M, Eichler EE. Analysis of segmental duplications and genome assembly in the mouse. Genome Res. 2004; 14:789–801. [PubMed: 15123579]

63. Bailey JA, Eichler EE. Primate segmental duplications: crucibles of evolution, diversity and disease. Nat Rev Genet. 2006; 7:552–64. [PubMed: 16770338]

64. Betran E, Wang W, Jin L, Long MY. Evolution of the Phosphoglycerate mutase processed gene in human and chimpanzee revealing the origin of a new primate gene. Molecular Biology and Evolution. 2002; 19:654–663. [PubMed: 11961099]

65. Rosso L, et al. Birth and rapid subcellular adaptation of a hominoid-specific CDC14 protein. PLoS Biol. 2008; 6:e140. [PubMed: 18547142]

66. Yu HJ, et al. Origination and evolution of a human-specific transmembrane protein gene, c1orf37-dup. Human Molecular Genetics. 2006; 15:1870–1875. [PubMed: 16644869]

67. Rosso L, Marques AC, Reichert AS, Kaessmann H. Mitochondrial targeting adaptation of the hominoid-specific glutamate dehydrogenase driven by positive Darwinian selection. PLoS Genet. 2008; 4:e1000150. [PubMed: 18688271]

68. Marques AC, Vinckenbosh N, Brawand D, Kaessmann H. Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. Genome Biology. 2008; 9

69. Smith, EL. The Enzymes. Boyer, PD., editor. Academic Press; New York: 1975. p. 293-367.

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. 2005; 79:65–73. [PubMed: 15578726]

71. Byun-McKay SA, Geeta R. Protein subcellular relocalization: a new perspective on the origin of novel genes. Trends in Ecology & Evolution. 2007; 22:338–344. [PubMed: 17507112]

72. Brennan G, Kozyrev Y, Hu SL. TRIMCyp expression in Old World primates Macaca nemestrina and Macaca fascicularis. Proc Natl Acad Sci U S A. 2008; 105:3569–74. [PubMed: 18287033]

73. Virgen CA, Kratovac Z, Bieniasz PD, Hatziioannou T. Independent genesis of chimeric TRIM5-cyclophilin proteins in two primate species. Proc Natl Acad Sci U S A. 2008; 105:3563–8. [PubMed: 18287034]

74. Wilson SJ, et al. Independent evolution of an antiviral TRIMCyp in rhesus macaques. Proc Natl Acad Sci U S A. 2008; 105:3557–62. [PubMed: 18287035]

75. Akiva P, et al. Transcription-mediated gene fusion in the human genome. Genome Res. 2006; 16:30–6. [PubMed: 16344562]

76. Babushok DV, et al. A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids. Genome Research. 2007; 17:1129–1138. [PubMed: 17623810]

77. Parra G, et al. Tandem chimerism as a means to increase protein complexity in the human genome. Genome Res. 2006; 16:37–44. [PubMed: 16344564]

78. Kleene KC. A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. Mechanisms of Development. 2001; 106:3–23. [PubMed: 11472831]

79. She X, et al. The structure and evolution of centromeric transition regions within the human genome. Nature. 2004; 430:857–64. [PubMed: 15318213]

80. Fontanillas P, Hartl DL, Reuter M. Genome organization and gene expression shape the transposable element distribution in the Drosophila melanogaster euchromatin. PLoS Genet. 2007; 3:e210. [PubMed: 18081425]

81. Bai Y, Casola C, Betran E. Evolutionary origin of regulatory regions of retrogenes in Drosophila. BMC Genomics. 2008; 9:241. [PubMed: 18498650]

82. Wang PJ. X chromosomes, retrogenes and their role in male reproduction. Trends in Endocrinology and Metabolism. 2004; 15:79–83. [PubMed: 15036254]

83. Shiao MS, et al. Origins of new male germ-line functions from X-derived autosomal retrogenes in the mouse. Molecular Biology and Evolution. 2007; 24:2242–2253. [PubMed: 17646254]

84. Sturgill D, Zhang Y, Parisi M, Oliver B. Demasculinization of X chromosomes in the Drosophila genus. Nature. 2007; 450:238–41. [PubMed: 17994090]

85. Hense W, Baines JF, Parsch J. X chromosome inactivation during Drosophila spermatogenesis. PLoS Biol. 2007; 5:e273. [PubMed: 17927450]

86. Lahn BT, Page DC. Four evolutionary strata on the human X chromosome. Science. 1999; 286:964–7. [PubMed: 10542153]

87. Skaletsky H, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature. 2003; 423:825–37. [PubMed: 12815422]

88. McLysaght A. Evolutionary steps of sex chromosomes are reflected in retrogenes. Trends Genet. 2008

89. Veyrunes F, et al. Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. Genome Res. 2008; 18:965–973. [PubMed: 18463302]

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. 2007; 8:R243. [PubMed: 18021405]

91. Ross MT, et al. The DNA sequence of the human X chromosome. Nature. 2005; 434:325–37. [PubMed: 15772651]

92. Betran E, Emerson JJ, Kaessmann H, Long M. Sex chromosomes and male functions - Where do new genes go? Cell Cycle. 2004; 3:873–875. [PubMed: 15190200]

93. Roy SW, Gilbert W. The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet. 2006; 7:211–21. [PubMed: 16485020]

94. Coulombe-Huntington J, Majewski J. Characterization of intron loss events in mammals. Genome Res. 2007; 17:23–32. [PubMed: 17108319]

95. Fink GR. Pseudogenes in yeast? Cell. 1987; 49:5–6. [PubMed: 3549000]

96. Goffeau A, et al. Life with 6000 genes. Science. 1996; 274(563):546–7. [PubMed: 8849441]

97. Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. Splicing and the evolution of proteins in mammals. Plos Biology. 2007; 5:343–353.

98. Kaiser J. DNA sequencing. A plan to capture human diversity in 1000 genomes. Science. 2008; 319:395. [PubMed: 18218868]

99. Tam OH, et al. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature. 2008

100. Watanabe T, et al. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature. 2008; 453:539–43. [PubMed: 18404146]

101. Wang PJ, Page DC. Functional substitution for TAF(II)250 by a retroposed homolog that is expressed in human spermatogenesis. Human Molecular Genetics. 2002; 11:2341–2346. [PubMed: 12217962]

102. Nielsen R, et al. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 2005; 3:e170. [PubMed: 15869325]

103. Jones CD, Begun DJ. Parallel evolution of chimeric fusion genes. Proc Natl Acad Sci U S A. 2005; 102:11373–8. [PubMed: 16076957]

104. 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. Molecular Biology and Evolution. 2007; 24:732–742. [PubMed: 17179138]

105. Pain D, Chirn GW, Strassel C, Kemp DM. Multiple retropseudogenes from pluripotent cell-specific gene expression indicates a potential signature for novel gene identification. J Biol Chem. 2005; 280:6265–8. [PubMed: 15640145]

106. Huang YT, Chen FC, Chen CJ, Chen HL, Chuang TJ. Identification and analysis of ancestral hominoid transcriptome inferred from cross-species transcript and processed pseudogene comparisons. Genome Res. 2008; 18:1163–70. [PubMed: 18369177]

Box 1

Retrocopies as genomic archives

Generally, retrocopies may serve as useful genomic markers of transcript activity during evolution. For example (as indicated in the 'Rate of retroposition' section), as retroposition is mediated by LINE elements, the rate of retrocopy generation (which may be calculated on the basis of the divergence of retrocopies and parental genes at 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 germline and/or the early embryo, the number of retrocopies should reflect parental gene activity during these stages[11,12]. Consistently, well-known housekeeping genes and/or genes with high germline/early embryo expression levels have produced many retrocopies[11,12,105]. Thus, retrocopies could serve as unique markers to shed light on the tissue origin of retroposition by correlating parental gene expression during different male/female germline or early embryonic stages with the abundance of their retrocopy offspring in the genome. The better the correlation observed in such an analysis, the more retrocopies would have emerged in a given germline/embryonic cell type.

Finally, the fact that retrocopies reflect their parental transcript structures have been exploited to detect previously unannotated or extinct, "fossil" transcripts[25,106]. For example, in a recent study, the authors reconstructed ancestral transcripts present in the common human-chimpanzee ancestor based on retrocopy sequences and inferred potential exon gains and losses in humans/chimpanzees based on their analysis[106].

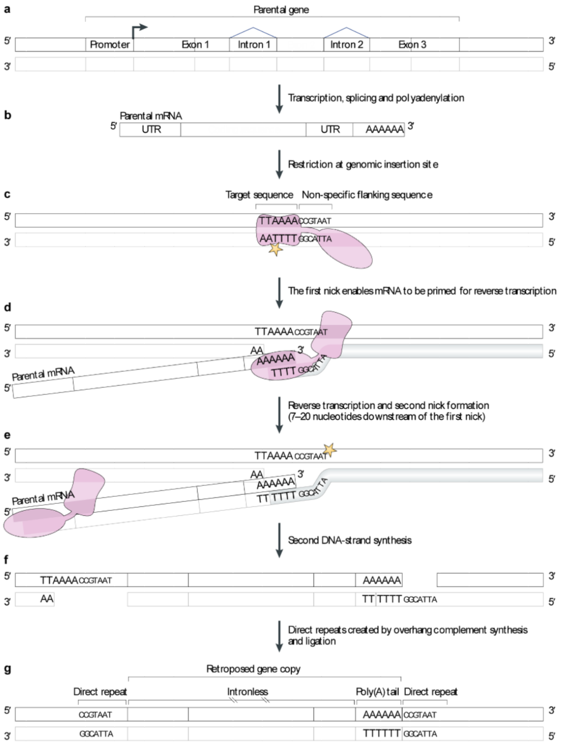

Figure 1. Mechanism of gene retroposition

(A) Gene retroposition is initiated with the transcription of a parental gene by RNA polymerase II and (B) further processing of its RNA (splicing and polyadenylation), which produces a mature mRNA. (C) Gene retroposition is mediated by the L1 endonuclease domain (pink hourglass) that creates a first nick (yellow star) at the genomic site of insertion at the TTAAAA target sequence. (D) This nick enables the priming of the reverse transcription (by the L1 reverse transcription domain; pink oval shape), which uses the parental mRNA as template. (E) Second strand nick generation (precise mechanism not known). (F) Second DNA strand synthesis (precise mechanism not known). (G) Complementary DNA synthesis in overhang regions created by the two nicks, which creates a duplication of the sequence flanking the target sequence, which is one of the molecular signatures of gene retroposition, in addition to the lack of introns and the presence of a poly-A tail (the direct repeats and the poly-A tail degenerate upon time and are therefore usually only detectable in recent retrocopies). The illustration is based on findings described in references [26-28].

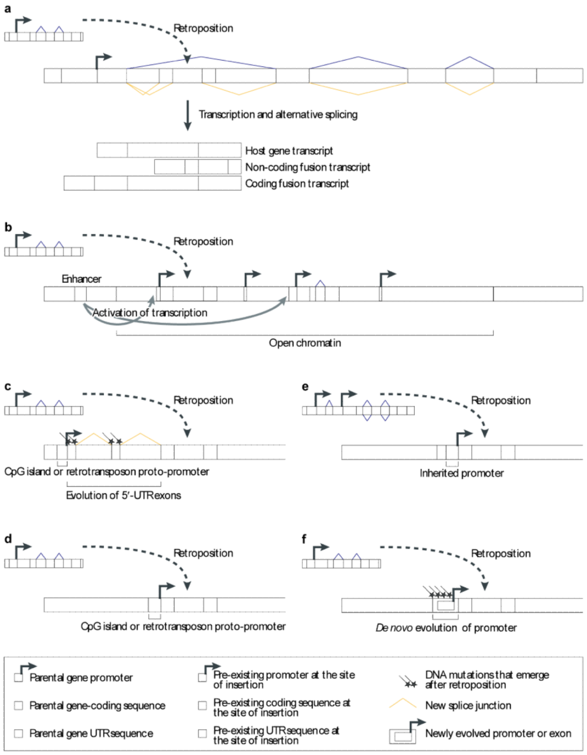
a
Retroposition
Transcription and alternative splicing
Host gene transcript
Non-coding fusion transcript
Coding fusion transcript
b
Retroposition
Enhancer
Activation of transcription
Open chromatin
c
Retroposition
CpG island or retrotransposon proto-promoter
Evolution of 5′-UTR exons
e
Retroposition
Inherited promoter
d
Retroposition
CpG island or retrotransposon proto-promoter
f
Retroposition
De novo evolution of promoter
Parental gene promoter
Parental gene-coding sequence
Parental gene UTR sequence
Pre-existing promoter at the site of insertion
Pre-existing coding sequence at the site of insertion
Pre-existing UTR sequence at the site of insertion
DNA mutations that emerge after retroposition
New splice junction
Newly evolved promoter or exon

Figure 2. Source of retrogene promoters

The figure illustrates various scenarios that lead to the transcription of retroposed gene copies. (A) Retrocopies may insert into intronic sequences of host genes. The evolution and/or presence of splicing signals enable 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 (where the entire open reading frame 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, due to the increased accessibility for the transcriptional machinery. The presence of enhancer elements from neighboring 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 neighborhood via the acquisition of a new untranslated exon/intron structure. (D) Recruitment of promoters from retrotransposons or CpG proto-promoters. (E) Inheritance of parental promoters through alternative transcriptional start site usage of the parental gene. (F) *De novo* promoter evolution in the 5′ flanking region of the insertion site by single nucleotide substitutions.

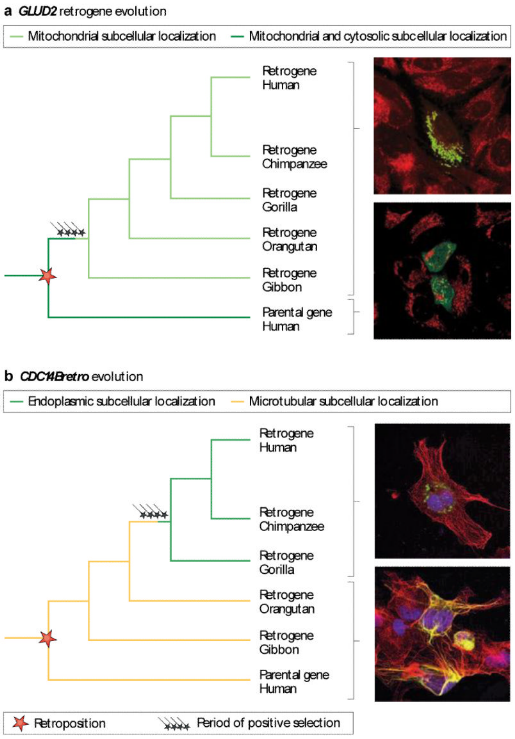

Figure 3. Subcellular adaptation of proteins encoded by new duplicate genes

(A) Illustration of 2 scenarios for the evolution of duplicated genes (red and green) and their products. Each gene and its encoded protein are represented with one color. Distinct protein shapes indicate distinct functions. Three different protein localizations (cytosolic, endoplasmic reticulum, or secreted proteins) are indicated in a schematic cell. Positively selected substitutions responsible for subcellular changes or changes in protein function are indicated (arrows). See main text for references and further details. (B) Adaptive evolution of two primate specific retrogenes (*GLUD2* left, *CDC14Bretro* right). Phylogenetic trees indicate retroduplication events. Periods of adaptive evolution and reconstructed subcellular localizations are indicated. Microscopy images display representative subcellular phenotypes for the indicated branches. Markers on the left: protein localization (green), nuclear DNA (blue), and microtubules (red). Yellow signals indicate an overlap of the protein with microtubules. Markers on the right: protein localization (green) and mitochondria (red).

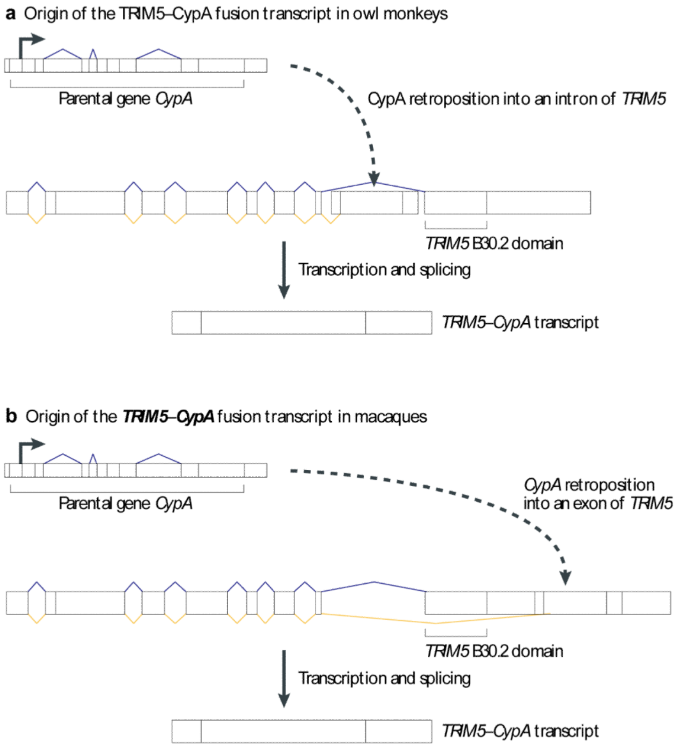
a Origin of the TRIM5–CypA fusion transcript in owl monkeys
Parental gene CypA
CypA retroposition into an intron of TRIM5
TRIM5 B30.2 domain
Transcription and splicing
TRIM5–CypA transcript
b Origin of the TRIM5–CypA fusion transcript in macaques
Parental gene CypA
CypA retroposition into an exon of TRIM5
TRIM5 B30.2 domain
Transcription and splicing
TRIM5–CypA transcript

Figure 4. Origin of TRIM5-CypA gene fusions in macques and owl monkeys

(A) Retroposition of *CypA* into an intron of the *TRIM5* gene from macaques and the resulting fusion gene is shown (similar to the process displayed in Fig. 2A). (B) An independent retroposition of *CypA* into the UTR of *TRIM5* in owl monkeys is shown, also resulting in a new *TRIM5-CypA* fusion gene. Please refer to Fig. 2 for the colour code and to the main text for details.

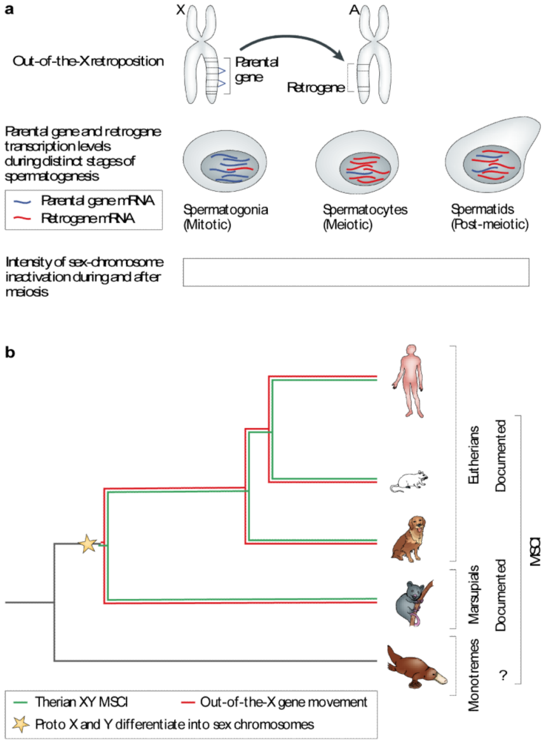
a
X
A
Out-of-the-X retroposition
Parental gene
Retrogene
Parental gene and retrogene transcription levels during distinct stages of spermatogenesis
Parental gene mRNA
Retrogene mRNA
Spermatogonia (Mitotic)
Spermatocytes (Meiotic)
Spermatids (Post-meiotic)
Intensity of sex-chromosome inactivation during and after meiosis
b
Eutherians
Documented
Marsupials
Documented
Monotremes
?
MSCI
Therian XY MSCI
Out-of-the-X gene movement
Proto X and Y differentiate into sex chromosomes

Figure 5. Retrogenes, MSCI, and the emergence of mammalian sex chromosomes

(A, upper part) Illustration of the retroposition of an X-linked parental gene to an autosome. (A, lower part) Illustration of the expression of X-linked parental genes and their autosomal retrogene copies before (in spermatogonial cells), during (spermatocytes), and after (spermatids) the process of meiotic sex chromosome inactivation (MSCI). (B) The evolutionary onset for the selectively driven out of X retroduplication process and MSCI, as well as the inferred origin of therian (eutherians/placental mammals and metatherians/marsupials) sex chromosomes. See main text for further explanations.

Table 1

Representative retrogenes in mammals and fruitflies.

Genes	Phylogenetic distribution	Features (Chromosomal origin / structure / type of selection / function)	References
Primates			
GLUD2	Hominoids	Into X, positive selection, subcellular adaptation, adaptation to (neurotransmitter) glutamate metabolism	23,67
CDC14Bretro	Hominoids	Positive selection, subcellular adaptation, derived from cell cycle gene, brain/testis-specific expression	37,65
c1orf37-dup	Humans	Positive selection, transmembrane protein	66
PGAM3	Old World primates	Positive selection, phosphoglycerate mutase	64
TRIM5-CypA gene	Macaque lineage	Chimeric gene, retrovirus restriction, CypA portion derives from retroposition	72-74
TRIM5-CypA gene	New World monkeys	Chimeric gene, retrovirus restriction, CypA portion derives from retroposition	20
PIP5K1A-PSMD4 retrogene		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	37,101
RBMXL1	Old World primates	X-derived, chimeric gene, fusion to host gene UTR	
Utp14c	Primates	X-derived, chimeric gene, evidence for it to be 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 which drives its transcription	51
Mammals			
Cstf2t	All Mammals	X-derived, chimeric gene, required for male fertility, crucial for proper polyadenylation in meiosis/post-meiosis	43
HNRNPGT	Therians	X-derived, required for male fertility	44
Pgk2	Eutherians	X-derived, promoter inherited from parent, acquisition of a testisspecific 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 X, among the 5 most positively selected gene in human-chimp comparison	102
Drosophila			
jingwei (jgw)	*D. yakuba, santomea* and *teisseri*	Chimeric gene, positive selection, retrocopy encoded ADH domain evolved new substrate (alcohol) specificity	21,48
Sphinx(spx)	*D. melanogaster*	Chimeric gene, positive selection, retrocopy evolved into non-coding RNA gene that promotes male-female courtship	24,49
Adh-Twain	*D. subobscura, guanche* and *madeirensis*	Chimeric gene, positive selection, putative functional adaptation to new substrate specificity	103
mojoless (mjl)	*Drosophila genus*	X-derived, required for male fertility	104
Dntf-2r	*D. melanogaster subgroup*	Substitutions in an upstream proto-promoter element appear to have provided this gene with a new, testis-specific promoter	

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 may obtain. We refer to most of these genes in the main text.

*Identified in mouse, phylogenetic distribution not established.