Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press

# Origins, evolution and phenotypic impact of new genes

Henrik Kaessmann

*Genome Res.* published online July 22, 2010
Access the most recent version at doi:10.1101/gr.101386.109

| | |
|---|---|
| **P<P** | Published online July 22, 2010 in advance of the print journal. |
| **Accepted Manuscript** | Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version. |
| **License** | |
| **Email Alerting Service** | Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.** |

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
**https://genome.cshlp.org/subscriptions**

Copyright © 2010, Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



# Origins, evolution and phenotypic impact of new genes

Running title: Evolution of new genes

Henrik Kaessmann

Center for Integrative Genomics, University of Lausanne, Genopode Building, CH-1015 Lausanne, Switzerland.

E-mail: Henrik.Kaessmann@unil.ch

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



## Abstract

**Ever since the pre-molecular era, the birth of new genes with novel functions has been considered to be a major contributor to adaptive evolutionary innovation. Here, I review the origin and evolution of new genes and their functions in eukaryotes, an area of research that has made rapid progress in the past decade thanks to the genomics revolution. Indeed, recent work has provided initial whole-genome views of the different types of new genes for a large number of different organisms. The array of mechanisms underlying the origin of new genes is compelling, extending way beyond the traditionally well-studied source of gene duplication. Thus, it was shown that novel genes also regularly arose from messenger RNAs of ancestral genes, protein-coding genes metamorphosed into new RNA genes, genomic parasites were co-opted as new genes and that both protein and RNA genes were composed from scratch (i.e., from previously nonfunctional sequences). These mechanisms then also contributed to the formation of numerous novel chimeric gene structures. Detailed functional investigations uncovered different evolutionary pathways that led to the emergence of novel functions from these newly minted sequences and, with respect to animals, attributed a potentially important role to one specific tissue – the testis – in the process of gene birth. Remarkably, these studies also demonstrated that novel genes of the various types significantly impacted the evolution of cellular, physiological, morphological, behavioral and reproductive phenotypic traits. Consequently, it is now firmly established that new genes have indeed been major contributors to the origin of adaptive evolutionary novelties.**

## Introduction

What is the nature of mutations underlying adaptive evolutionary innovations? In addition to subtle genetic modifications of preexisting ancestral genes that can lead to differences in their (protein or RNA) sequences or activities, new genes with novel functions may have significantly contributed to the evolution of lineage- or species-specific phenotypic traits. Consequently, the process of the "birth" and evolution of novel genes has attracted much attention from biologists in the past. Indeed, quite remarkably, considerations pertaining to the origin and functional fate of new genes trace back to a time when the molecular nature of genes had not yet been established. Based on cytological observations of chromosomal duplications, Haldane and Muller already hypothesized in the 1930s that new gene functions may emerge from refashioned copies of old genes (Haldane 1933; Muller 1935), highlighting for the first time the potential importance of gene duplication for the process of new gene origination. The early notions that gene duplication provides a significant reservoir for the emergence of genes and hence phenotypic adaptation have now been globally confirmed (but also refined) based on numerous large- and small-scale molecular studies that were facilitated by the genomics revolution. New duplicate genes have been shown to be abundant in all eukaryotic genomes sequenced to date and to have evolved pivotal functional roles (Lynch 2007).

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



However, studies from the genomics era have also accelerated the discovery of fascinating novel mechanisms underlying the emergence of new genes. These include the origin of new protein-coding and RNA genes "from scratch" (that is, from previously nonfunctional genomic sequences), various types of gene fusions, and the formation of new genes from RNA intermediates. It is now well established that all of these mechanisms have signicantly contributed to functional genome evolution and phenotypic change, which further underscores the importance of novel genes for organismal evolution.

In this review, I discuss in detail the different genomic sources of new genes in eukaryotes (with a particular emphasis on animals) and assess their relative contributions and functional implications in different species and evolutionary lineages. I also examine how new protein or RNA functions may evolve from newly minted gene structures and discuss the associated selective forces. I then discuss a hypothesis that suggests a key role of one tissue – the testis – in the establishment of new functional genes. Finally, I highlight recent new developments in the field and identify potential future research directions. Notably, I focus on recent developments in this review, while referring to previous reviews and other literature for details pertaining to long established concepts and earlier findings.

## Gene duplication – raw material for the emergence of new genes

Gene duplication is a very common phenomenon in all eukaryotic organisms (but also in prokaryotes – reviewed in (Romero and Palacios 1997)) that may occur in several different ways (Lynch 2007). Traditionally, DNA-mediated duplication mechanisms have been considered and widely studied in this context, although peculiar intronless duplicate gene copies may also arise from RNA sources (see further below). DNA duplication mechanisms include small-scale events, such as the duplication of chromosomal segments containing whole genes or gene fragments (termed segmental duplication), which are essentially outcomes of misguided recombination processes during meiosis (Fig. 1A). However, they also include duplication of whole genomes through various polyploidization mechanisms (Lynch 2007; Conant and Wolfe 2008; Van de Peer et al. 2009). Thus, duplicate gene copies can arise in many different ways. But what is their functional fate and evolutionary relevance?

***Gene duplication and new gene functions.*** At least since a famous monograph, authored by Susumu Ohno, was published over 40 years ago (Ohno 1970), the word has spread that gene duplication may underlie the origin of many or even most novel genes and hence represents an important process for functional innovation during evolution. Essentially and consistent with earlier ideas (Haldane 1933; Muller 1935), Ohno emphasized that the presence of a second copy of a gene would open up unique new opportunities in evolution by allowing one of the

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



two duplicate gene copies to evolve new functional properties, whereas the other copy is preserved to take care of the ancestral (usually important) function (the concept of neofunctionalization). Ohno also reviewed that duplicate genes can be preserved by natural selection for gene dosage, thus allowing an increased production of the ancestral gene product (Ohno 1970). Finally, it should emphasized that it has been widely agreed for a long time that the most probable fate of a duplicate gene copy is pseudogenization (Ohno 1972) and that hence the majority of duplicate gene copies are eventually lost from the genome.

While these fundamental hypotheses have been confirmed by a large body of data, they have since also been significantly extended and refined. In particular, in addition to the process of neofunctionalization (i.e., the emergence of new functions from one copy - Ohno's basic concept), it was proposed that the potentially multiple functions of an ancestral gene may be partitioned between the two daughter copies. This process was dubbed "subfunctionalization" and may be shaped by natural selection or involve purely neutral processes (Force et al. 1999; Conant and Wolfe 2008; Innan and Kondrashov 2010).

Global genomic screens combined with detailed experimental scrutiny have uncovered numerous intriguing examples for each of these models in many organisms, solidly supporting their validity. Detailed analyses of young duplicate genes have been particularly informative, because many of the details associated with the emergence of new genes from gene duplicates become obscured over longer periods of time (Long et al. 2003). A particularly illustrative case of neofunctionalization, arguably the most intriguing fate of a duplicate gene, occurred in the course of the recent duplication of a pancreatic ribonuclease gene in leaf-eating monkeys. Zhang et al. demonstrated that after duplication in an African leaf-eating monkey, the protein encoded by one of the copies of the ancestral *RNASE1* gene rapidly adapted at specific sites to derive nutrients from bacteria in the foregut under the influence of strong positive selection (Zhang et al. 2002). Remarkably, both the duplication and subsequent adaptation of this gene was later shown to have occurred independently in a very similar manner in an Asian leaf-eating monkey (Zhang 2006). Thus, these *RNASE1* duplications represent striking cases of convergent molecular evolution. They were likely facilitated by the frequent occurrence of segmental duplication, which allows similar duplication events that are highly beneficial to be repeatedly fixed during evolution. More generally, the convergent *RNASE1* duplications are in line with several other recent reports that include other cases of new gene formation (see below) and therefore lend further support to the more general idea that adaptive genome evolution is, to some extent, predictable (Stern and Orgogozo 2009). Numerous other classical or recent examples from diverse organisms could be discussed here that illustrate the immense potential that DNA-based gene duplication has held for phenotypic evolution in different organisms (reviewed in, e.g., (Li 1997; Long et al. 2003; Zhang 2003; Lynch 2007; Conant and Wolfe 2008)).

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



***Duplication of non-coding RNAs.*** Suffice it to add in this review that studies pertaining to the origin of novel genes from duplicated DNA segments have begun to be extended beyond the traditionally studied protein-coding genes, thanks to the rapid recent advances in the genomics field. For example, it has become clear that microRNAs (miRNAs), small RNA molecules that have emerged as major post-transcriptional regulators (Carthew and Sontheimer 2009), have expanded and functionally diversified during evolution by gene duplication (Hertel et al. 2006). Interestingly, several individual studies indicate that the X chromosome may provide a particularly fruitful ground for the origination of new lineage-specific miRNAs (Zhang et al. 2007; Devor and Samollow 2008; Murchison et al. 2008; Guo et al. 2009), a pattern that may be explained by the specific sex-related forces that have shaped the X, given that new X-born miRNAs appear to be predominantly expressed in male-reproductive tissues. Segmental gene duplication also seems to play a major role for the expansion of another class of small RNAs, Piwi-interacting RNAs (piRNAs, (Malone and Hannon 2009)), which are expressed in the germline and are thought to be mainly involved in transposon control. A recent study revealed that piRNA clusters rapidly expanded through segmental duplication in primate and rodent genomes, a process driven by intense positive selection (Assis and Kondrashov 2009). Segmental duplication therefore provides an efficient vehicle for the expansion of piRNA repertoires and hence allows organisms to swiftly evolve protection barriers against the lineage-specific expansion of transposable elements. There is so far little evidence for duplication of sequences transcribed into long noncoding RNAs (lncRNAs), an abundant class of non-translated RNAs (> 200 nucleotides in length), whose functional impact is only beginning to be understood (Mercer et al. 2009; Ponting et al. 2009). The paucity of known duplicated lncRNA genes is perhaps mainly due to their rapid sequence divergence, which may render the detection of such events difficult. Future work, which will benefit from the rapidly accumulating genomic and transcriptomic data, will clarify the role of gene duplication in the evolution of new lncRNA genes with altered or novel functions.

***Global patterns.*** In spite of the numerous well-founded examples of functionally important newly minted genes that arose from duplicate gene copies, a more global picture of the functional relevance and adaptive value of the large number of duplicate gene copies scattered in genomes is only beginning to emerge. Only for some whole-genome duplication (WGD) events in model organisms (in particular yeast), global assessments of the relevance of duplicate genes for the emergence of new gene functions have been attempted (Conant and Wolfe 2008). However, WGD represents a special case of gene duplication, which involves specific selective pressures related to dosage balance of gene products that seem to significantly influence the fate of resulting gene duplicates. And even in the case of WGD, it remains largely unclear whether gene duplications often conferred novel functions or not (Conant and Wolfe 2008).

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



Thus, a more global understanding of the implications of gene duplication for the emergence of new gene functions and its importance relative to other mutational mechanisms that affect preexisting genes will have to await future efforts. However, a closer examination of the reported general distributions and characteristics of gene duplicates in different genomes is nevertheless instructive.

For example, analyses of fully sequenced genomes have revealed high rates of origin but also loss of duplicate genes (Lynch and Conery 2003; Demuth and Hahn 2009). New duplicates are estimated to be "born" at the rate of approximately 0.001-0.01 per gene per million years in eukaryotes (Lynch and Conery 2003; Lynch 2007), while the death rate of duplicates is at least an order of magnitude higher, consistent with the early notion (see above) that the fate of most duplicates is pseudogenization (Ohno 1972). Notably, not all functional categories of genes are equally prone to expand by duplication. In particular, a relatively small number of gene families (1.6-3%) with functions in, for example, immunity, host defense, chemosensation and reproduction show rapid, selectively driven copy number changes in various eukaryotic lineages, thus significantly contributing to their adaptive evolution (Emes et al. 2003; Demuth and Hahn 2009).

However, in addition to these commonalities, detailed whole-genome investigations also suggest intriguing fundamental differences with respect to the generation and functional fate of duplicates in different evolutionary lineages. For example, careful analyses in primates revealed a burst of segmental gene duplication in hominoids (humans and apes), especially in humans and the African apes (Marques-Bonet and Eichler 2009). Notably, many of these duplicates are dispersed and mediate major genomic rearrangements associated with disease. The accelerated fixation rate of segmental duplicons in hominoids could, in principle, be explained by the selective benefit of newly formed genes embedded within these regions, which outweigh deleterious effects in many cases (Marques-Bonet et al. 2009b). New gene formation in hominoids indeed seems to have profited from the substantial raw material provided by massive segmental duplication ((Marques-Bonet et al. 2009b), see below). However, the overall accelerated fixation rate of segmental duplicons in humans and apes is probably best explained by the reduction of the effective population size in the hominoid lineage. This reduction increased genetic drift and, at the same time, rendered purifying selection less efficient, thus probably allowing disproportionately high numbers of slightly deleterious segmental duplications to be fixed in hominoids compared to other species with larger long-term effective population sizes (and hence more efficient selection). This hypothesis is consistent with other types of molecular evolutionary data (Keightley et al. 2005; Gherman et al. 2007).

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



In addition to lineage-specific selection intensities, differences pertaining to the mutational basis of gene duplication can lead to different characteristics of segmental duplications between species. A good example is the finding that, in contrast to humans, recently duplicated chromosomal regions in the mouse are depleted in genes and transcripts (She et al. 2008). Detailed analyses suggest that species-specific distributions of retrotransposons, which represent major promoters of segmental duplication events (Marques-Bonet et al. 2009a), account for much of this discrepancy.

## RNA-based duplication and the emergence of "stripped-down" new genes

As outlined above, the traditionally studied DNA-mediated gene duplication mechanisms have significantly contributed to functional genome evolution and have provided many fundamental insights regarding new gene origination. However, new gene copies can also arise through an alternative, less well known duplication mechanism termed retroposition or retroduplication (Brosius 1991; Long et al. 2003; Kaessmann et al. 2009). In this mechanism, a mature messenger RNA (mRNA) that is transcribed from a "parental" source gene is reverse transcribed into a complementary DNA copy, which is then inserted into the genome (Fig. 1B). The enzymes necessary for retroposition (in particular the reverse transcriptase), is encoded by different retrotransposable elements in different species. In mammals, LINE-1 retrotransposons provide the required enzymatic machinery (Mathias et al. 1991; Feng et al. 1996; Esnault et al. 2000). Given that the resulting intronless retroposed gene copies (retrocopies) only contain the parental exon information (i.e., they usually lack parental introns and core promoter sequences), retrocopies were long thought to be consigned to the scrapheap of genome evolution and were routinely labeled as "processed pseudogenes" (Mighell et al. 2000). However, after anecdotal findings of individual functional retrocopies (so-called retrogenes) in the 1980s and 1990s, a surprising number of retrogenes could be discovered with the advent of the genomics era. Notably, detailed analyses of this stripped-down type of new genes have revealed previously unknown mechanisms underlying the appearance of new genes and their functions and demonstrated that new retrogenes have contributed to the appearance of lineage-specific phenotypic innovations (Kaessmann et al. 2009).

***Sources of regulatory elements.*** The observation of numerous functional retrogenes in various genomes (detailed below) immediately raises the question of how retrocopies can obtain regulatory sequences that allow them to become transcribed – a precondition for gene functionality. Studies that sought to address this question uncovered various sources of retrogene promoters and regulators and therefore also provided general insights into how new genes can acquire promoters and evolve new expression patterns (Kaessmann et al. 2009). First, it was shown that the expression of new retrogenes often benefits from preexisting regulatory machinery and expression capacities of genes in their vicinity. Thus, retrogenes profited from the open chromatin state and accessory regulators

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



(enhancers/silencers) of nearby genes, directly fused to host genes into which they inserted (see also below), or captured bi-directional promoters of genes in their proximity (Vinckenbosch et al. 2006; Fablet et al. 2009; Kaessmann et al. 2009). Second, retrogenes recruited CpG dinucleotide-enriched proto-promoter sequences in their genomic vicinity not previously associated with other genes for their transcription (Fablet et al. 2009). Third, retrotransposons upstream of retrocopy insertions sites were shown to have provided retrogenes with regulatory potential (Zaiss and Kloetzel 1999; Fablet et al. 2009). Fourth, unexpectedly, retrogenes also seem to frequently have directly inherited alternative promoters embedded in parental transcripts that gave rise to them (Okamura and Nakai 2008; Kaessmann et al. 2009). Finally, basic retrogene promoters may sometimes have evolved *de novo* through small substitutional changes under the influence of natural selection (Betran and Long 2003; Bai et al. 2007). Remarkably, the process of promoter acquisition sometimes involved the evolution of new 5' untranslated exon-intron structures, which span the often substantial distances between the recruited promoters and retrogene insertion sites (Fablet et al. 2009).

***New retrogene functions.*** Given that retrocopies usually need to acquire regulatory elements for their transcription, retrocopies that eventually do become transcribed – a surprisingly frequent event (Vinckenbosch et al. 2006) – are much more prone to evolve novel functions (and less likely to be redundant) than gene copies arising from DNA-based duplication mechanisms. Indeed, a number of new retrogenes with intriguing functions have been identified. Detailed analyses of these retrogenes uncovered novel mechanisms underlying the emergence of new gene functions. For example, analyses of young retrogenes in primates not only revealed that retrogenes have contributed to hominoid brain evolution, but also identified different molecular levels at which new genes may adapt to new functions. Namely, in addition to evolving new spatial expression patterns relative to the parental source genes, the proteins encoded by these retrogenes evolved new biochemical properties (Burki and Kaessmann 2004) and/or subcellular localization patterns (Burki and Kaessmann 2004; Rosso et al. 2008a; Rosso et al. 2008b). The latter process, dubbed subcellular adaptation or relocalization, could be established and generalized as a new trajectory for the evolution of new gene functions after these observations (Marques et al. 2008; Kaessmann et al. 2009).

Other interesting retrogenes have recently been unveiled that exemplify the sometimes unexpected and curious pathways of evolutionary change. An example is a mouse retrocopy of a ribosomal protein gene (*Rps23*), of which there are hundreds in mammalian genomes and that usually represent nonfunctional retropseudogenes, consistent with the idea that duplication of these genes is usually redundant and/or is subject to dosage balance constraints. Yet the *Rps23* retrocopy evolved a completely new function, not by changes in the protein-coding sequence, but by being transcribed from the reverse strand and the incorporation of sequences flanking its

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



insertion site as new (coding and noncoding) exons (Zhang et al. 2009). This gave rise to a new protein (completely unrelated to that encoded by its parental gene), which had profound functional implications in that it conferred increased resistance in mice against the formation of Alzheimer-causing amyloid plaques.

Another intriguing recent case of new retrogene formation illustrates the far-reaching and immediate phenotypic consequences a retroduplication event may have. Parker et al. found that a retrocopy derived from a growth factor gene (*fgf4*) is solely responsible for the short-legged phenotype characteristic of several common dog breeds (Parker et al. 2009). Remarkably, the phenotypic impact of the *fgf4* retrogene seems to be a rather direct consequence of the gene dosage change associated with its emergence (i.e., increased FGF4 expression during bone development), given that its coding sequence is identical to that of its parental gene. The analysis of *fgf4* in dogs thus strikingly illustrates that gene duplication can immediately lead to phenotypic innovation (in this case a new morphological trait) merely through gene dosage alterations.

***Retrogenes and meiotic sex chromosome inactivation.*** Numerous other illuminating cases of retrogenes known to have evolved diverse functions in species ranging from primates and flies to plants have recently been described (reviewed in (Kaessmann et al. 2009)). However, global surveys of retroposition conducted in mammals and fruitflies have also identified a common theme uniting a significant subset of new retrogenes in these species: expression and functionality in testes. While these retrogenes seem to have evolved a variety of functional roles (a process that may have a mechanistic basis and was likely influenced by sexual selection – see below), the functions of a disproportionately high number among them are apparently associated with the transcriptional inactivation of the sex chromosomes in the male germline during and (to a lesser extent) after meiosis (Turner 2007). Thus, it now seems clear that the many mammalian retrogenes that stem from the X have been fixed during evolution and shaped by natural selection to compensate for the transcriptional silencing of their parental (often housekeeping) genes during male germline silencing of the X (Bradley et al. 2004; Rohozinski and Bishop 2004; Potrzebowski et al. 2008). Indeed, systematic analyses of chromosomal positions of parental genes and their daughter retrocopies revealed that a larger than expected number of autosomal retrogenes are derived from parental genes located on the X in various mammals (Emerson et al. 2004; Potrzebowski et al. 2008) and that these retrogenes are specifically expressed during and after meiosis, when their parental genes are silenced (Potrzebowski et al. 2008). Consequently, testis functions of parental genes can be considered to have spread or “moved” to the autosomes, a process that was facilitated by the fact that the retroposition process readily transfers genes between chromosomes (more readily so than segmental duplication, which often occurs on the same chromosome). Notably, recent work (Vibranovski et al. 2009) indicates that meiotic sex chromosome inactivation may also underlie the export of retrogenes from the X in *Drosophila* (Betran et al. 2002).

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



Knowing the functional basis for this so called "out of X" movement of genes then also allowed dating of the evolutionary onset of mammalian meiotic sex chromosome silencing through assessments of the age of X-derived retrogenes. This work revealed that not only the mechanisms of meiotic sex chromosome silencing but also the sex chromosomes themselves originated in the common ancestor of placental mammals and marsupials (i.e., after the divergence from lineage of egg-laying monotremes) and hence are younger than previously thought (Potrzebowski et al. 2008). Notably, tracing the evolutionary origin of individual X-derived retrogenes also identified striking cases of independent parallel exports of key housekeeping genes in eutherians and marsupials, which illustrates the strong selective pressures that drove genes out of the X upon the emergence of sex chromosomes. Curiously, a recent study revealed that the X chromosome not only exported many genes but also preferentially accumulated new retrogenes upon therian (eutherian and marsupial) sex chromosome differentiation, apparently owing to the emerging sex-related (potentially antagonistic) selective forces (Potrzebowski et al. 2010).

***Retroduplication in different evolutionary lineages.*** Together, these examples illustrate that new retrogenes have been conducive to the evolution of new genome functions and phenotypic innovation. However, it should be noted that retroposition has contributed to the evolution of different eukaryotic lineages to highly varying degrees, because of fundamental differences related to the machinery responsible for this process. For example, the rate of retroduplication has been overall high in therian mammals because of the high activity of L1 retrotransposons, which provide the enzymes (reverse-transcriptase and endonuclease) necessary for this process (Kaessmann et al. 2009). Thus, thousands of retrocopies and over one-hundred functional retrogenes have been identified in the human genome (Vinckenbosch et al. 2006). Fruitfly genomes have also been found to contain many functional retrogenes (Betran et al. 2002; Bai et al. 2007; Zhou et al. 2008). In contrast, genomes from monotreme mammals and birds only contain very few retrocopies and lack functional retrogenes, owing to the absence of retrotransposons that could provide the appropriate retroposition machinery (Hillier et al. 2004; Kaessmann et al. 2009). However, eukaryotic lineages previously thought to be depauperate in terms of retroposition activity, such as plants, have recently unveiled a surprisingly large number of apparently selectively constrained retrogenes (Wang et al. 2006; Zhu et al. 2009). Thus, retroduplication has contributed to the phenotypic evolution of many multicellular eukaryotes, ranging from mammals and insects to plants, by giving rise to many functional new genes, although this contribution has been more variable than that of the more common and widespread DNA-mediated duplication mechanisms.

## Formation of new gene structures by retrotransposon-mediated transduction

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



An alternative mode by which retrotransposons could contribute to the formation of new gene structures was identified in the late 1990s (Moran et al. 1999). The authors showed that, in addition to process of retroposition, in which the retrotransposons-derived enzymes generate copies of mature mRNAs (see section above), L1 retrotransposon transcripts can also directly carry downstream flanking genomic sequences with them. In this process, which is termed 3' transduction, the RNA transcription machinery reads through the weak retrotransposon polyadenylation signal and terminates transcription by using an alternative signal downsteam in the 3' flanking sequence (reviewed in (Cordaux and Batzer 2009)). Subsequent studies showed that many (~10%) of L1 and SVA retrotransposon insertions are associated with 3' transduction events, copying various genic elements into new genomic locations ((Cordaux and Batzer 2009) and references therein). An interesting recent study provided initial evidence that 3' transduction may have lead to the formation of new genes in primates (Xing et al. 2006). As part of a genome-wide analysis of SVA-mediated transduction, Xing et al. identified 143 events that transduced sequences of various sizes. Notably, three separate events transduced the entire *AMAC1L3* gene into 3 new genomic locations ~7-14 million years ago in the human/African ape ancestor. The novel gene copies were shown to be transcribed, but it was unclear whether they have been preserved by natural selection (Xing et al. 2006). Thus, while the functional relevance of this new gene family in African apes remains unclear, this study provides initial evidence that 3' transduction may represent yet another way by which retrotransposons have contributed to the functional evolution of the genome.

## Gene fusion – the origin of new chimeric genes

The process of gene fusion is defined as the fusion of two previously separate source genes into a single transcription unit – the so called fusion or chimeric gene (Long et al. 2003). Gene fusion is a fascinating mechanism of new gene origination that is almost bound to give rise to new functions given its combinatorial nature (assuming that the fusion gene is beneficial and selectively preserved). In agreement with this notion, a number of chimeric genes with important functions have been described (Long et al. 2003; Zhou and Wang 2008; Kaessmann et al. 2009). The various mechanisms underlying the formation of new chimeric gene structures and their evolutionary relevance are discussed in the following using representative examples.

***DNA-mediated gene fusions.*** A common theme underlying several of the different gene fusion mechanisms is gene duplication, which provides the necessary raw material for the emergence of new fusion genes, allowing ancestral gene functions to be preserved. Thus, chimeric genes often arise from juxtaposed pieces of duplicate gene copies through fission and fusion processes (Fig. 2A). For example, the dispersion and shuffling of numerous segmental gene copies in hominoids through various recombination and translocation events has led to the formation of many mosaic gene structures, some of which have become transcribed (Bailey et al. 2002; She et al. 2004;

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



Marques-Bonet et al. 2009a). Among these transcribed chimeras, there are several genes with known functions (e.g., the *Tre2* oncogene with testis expression; (Paulding et al. 2003)), or genes that have further expanded and show signatures of positive selection (e.g., *RANBP2*; (Ciccarelli et al. 2005)), suggesting that they evolved new beneficial functions. Juxtaposition of partial segmental duplicates also seems to rather frequently have led to the emergence of young functional genes in fruitflies, more often so than the apparently often redundant complete gene duplications (Zhou et al. 2008). These observations illustrate the evolutionary potential offered by DNA-based gene fusion events for the more recent evolution of animals. However, a number of highly modular ancient genes, sharing exons encoding specific protein domains, also attest to the functional importance of DNA-based exon shuffling (i.e., the exchange/fusion of individual exons) for early metazoan evolution (Patthy 1999).

Retroduplication is a mechanism that could be expected to lend itself well for the process of gene fusion, given that it readily moves gene sequences to new locations in the genome. Indeed, a number of functionally relevant fusion events involving retrogenes have been described. For example, retrocopies were shown to frequently have inserted into an intron of a host gene and to have become transcribed in the form of a fusion transcript together with host gene exons (Vinckenbosch et al. 2006; Kaessmann et al. 2009). Often, these retrocopies are transcribed with only 5'-untranslated exons of the host gene, as alternative splice variants, thus profiting from promoters from the host gene (see also above), while leaving host gene functions unaltered.

However, functional coding sequence fusions of host genes and retrogenes have occurred as well. A classical example is the testis-expressed *jingwei* (*jgw*) gene in *Drosophila* (Long and Langley 1993), the first young fusion gene described. This gene emerged through a series of events based on the fusion of parts of a segmental duplicate gene copy (*ynd*, which provided the regulatory elements) with a retrocopy of the alcohol dehydrogenase gene. Biochemical and evolutionary analysis further revealed that the *jgw*-encoded protein evolved a new functional role in hormone and pheromone metabolism under the influence of positive Darwinian selection (Zhang et al. 2004). Functionally important retrogene-host gene coding fusions have also occurred in mammals. Retrocopies from the *cyclophilin A* (*CypA*) gene, which encodes a protein that potently binds retroviral capsids, were shown to have integrated into the 3' end of the antiviral defense gene *TRIM5* in a New World monkey, replacing and functionally substituting the exons encoding the original capsid-binding domain from *TRIM5* (Sayah et al. 2004). Remarkably, a highly similar event was independently fixed in the Old World monkey lineage (Brennan et al. 2008), which illustrates the high selective benefit associated with the creation of this type of chimeric gene. Thus, the *TRIM5-CypA* gene fusions present striking cases of domain shuffling and, taken together, provide yet another fascinating example of convergent evolution in the field of new gene origination. With respect to the fusion of retrogenes with preexisting exons, it is finally noteworthy that this process seems to be rather prevalent

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



in plants (Wang et al. 2006; Zhu et al. 2009). While the functions and phenotypic implications of the majority of these plant chimeric genes remain to be explored, an interesting class of functional chimeric genes that involve fusions of mitochondrial retroposed gene copies and nuclear genes was identified in flowering plants (Nugent and Palmer 1991; Liu et al. 2009). Specifically, it was shown that mitochondrial genes became relocated to the nuclear genome, probably via RNA intermediates (Nugent and Palmer 1991), forming chimeras with preexisting nuclear genes. Notably, in many cases the ancestral nuclear genes provided targeting signals for import of the mitochondrion-derived protein back into mitochondria (Liu et al. 2009). Thus, this type of gene fusion readily allowed for transfer of mitochondrial genes into the nucleus while mitochondrial functions could be maintained.

***Transcription-mediated gene fusions.*** In addition to the genome-based juxtapositions and "permanent" fusions of genes or gene fragments described above, recent work uncovered an alternative gene fusion mechanism that combines exons from independent consecutive genes in the genome at the transcription level by intergenic splicing (Fig. 2B). Given that this mechanism draws from exons of preexisting genes, it does not represent a true process of new gene formation, but is nevertheless interesting to discuss here, given that it gives rise to new *transcription* units with potentially novel functions that may sometimes be fixed as new genes in the genome through secondary events (see below). Transcription-mediated gene fusion was long thought to be exceedingly rare, but after the discovery of individual cases early in the past decade (e.g., (Thomson et al. 2000)), genome-wide surveys unearthed large numbers of transcription-induced chimeras (Akiva et al. 2006; Parra et al. 2006; Denoeud et al. 2007). Notably, many of these chimeras involve fusions of protein-coding exons from adjacent genes. But although their expression levels are sometimes relatively high (Denoeud et al. 2007) and individual characterizations suggest specific subcellular localizations of encoded products with respect to the proteins encoded by the involved partner genes (Thomson et al. 2000; Pradet-Balade et al. 2002), the functional and evolutionary potential of these fused transcripts remains to be explored. Also, their evolutionary origin (presumably through the emergence and fixation of intergenic splice sites) and level of selective preservation between species have yet to be documented. Interestingly, however, at least one of the transcription-induced chimeric mRNAs was shown to have become fixed in the genome during evolution as a separate new gene through the process of retroposition ((Akiva et al. 2006); Fig. 2B). Babushok et al. showed that this new retrogene (termed *PIP5K1A*) emerged in the common hominoid ancestor, became specifically expressed in testes, experienced a phase of intense positive selection and shows significant affinity for cellular ubiquitinated proteins (reflecting a modified activity of one of the parental proteins), which suggests a new and beneficial functional role of the encoded protein in apes (Babushok et al. 2007).

## Gene origination from scratch

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



As noted above, the origin of new genes was long believed to be intimately linked to the process of gene duplication (Ohno 1970). Consistent with this notion (and as discussed in this review), new genes were usually found to be associated with duplicated genomic raw material in one way or another. Yet, what one would probably intuitively associate with true gene "birth" and what could, arguably, be considered the most intriguing mode (also because it is almost bound to provide a new function), is the emergence of new genes "from scratch". In other words, new genes arise from previously nonfunctional genomic sequence, unrelated to any preexisting genic material (Fig. 3).

***De novo emergence of protein-coding genes.*** The *de novo* origin of entire protein-coding genes was long considered to be highly unlikely. For instance, in agreement with his contemporary gene duplication advocates, François Jacob noted in an influential essay that the "probability that a functional protein would appear *de novo* by random association of amino acids is practically zero" and that therefore the "creation of entirely new nucleotide sequence could not be of any importance in the production of new information" (Jacob 1977).

In spite of these notions, recent work has uncovered a number of new protein-coding genes that apparently arose from previously noncoding (and non-repetitive) DNA sequences. Probably the first such case described in the literature is presented by the *morpheus* gene family that emerged in an Old World primate ancestor (Johnson et al. 2001). Although the details regarding the emergence of the original coding sequence remain unclear, the lack of any corresponding orthologous sequences outside of Old World primates suggest a *de novo* origin for this gene family. Notably, Johnson et al. revealed that the ancestor of this gene family massively expanded by segmental duplication in hominoids, and that the various morpheus gene copies show spectacular signatures of positive selection in their coding sequences, suggestive of exceedingly high rates of adaptive protein evolution (Johnson et al. 2001). Although the precise functional roles of the *morpheus* genes have not yet been determined, the strong selective pressures associated with their evolution suggest important and rapidly evolving functions of the encoded proteins in humans and apes.

Other studies have followed suit and have provided a more detailed picture of *de novo* gene origination. For example, 14 *de novo*-originated genes have been identified in *Drosophila* (Levine et al. 2006; Zhou et al. 2008), the majority of which are specifically expressed in testes. This suggests that *de novo* gene formation may have contributed an unexpectedly large proportion of new genes in this genus. Other studies have reported new genes that evolved *de novo* in yeast and primates (Cai et al. 2008; Knowles and McLysaght 2009; Toll-Riera et al. 2009). For example, Knowles and McLysaght recently identified three genes that seem to have arisen from scratch on the human lineage (Knowles and McLysaght 2009). Detailed analyses of these human-specific genes, which involved

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



comparisons with corresponding non-coding sequences from closely-related primate relatives, revealed that a few mutational events after the separation of the human and chimpanzee lineages abolished "disabling" nucleotides in the ancestral open reading frame precursors (Fig. 3), allowing relatively long coding sequences to emanate in humans. Importantly, the functionality of these new human genes is supported by evidence for translation of their coding sequences.

Together, these studies suggest that the *de novo* emergence of new protein-coding genes is more likely than previously thought, although more work is required to elucidate the functional relevance and potential phenotypic implications of the reported cases. More generally, the available studies illustrate the two key events that must precede the birth and fixation of a new protein-coding gene from ancestrally noncoding DNA region (Fig. 3): (i) The DNA must become transcriptionally active and (ii) it must also evolve a translatable open reading frame that encodes a potentially beneficial protein. The former may be readily achieved, given the high transcriptional activity of the genome and the various mechanisms that allow new genes to recruit regulatory sequences (see retrogene section above). A more global assessment of the probability for the latter will have to await future studies. These will also further our understanding of the evolutionary importance of *de novo* protein-coding gene birth relative to other mechanisms of new gene formation.

## Origins of noncoding RNA genes

Recent transcriptome studies have unveiled an unexpectedly rich repertoire of noncoding RNA species, which, in mammals, are derived from hundreds of small and thousands of lncRNA loci (Carthew and Sontheimer 2009; Ponting et al. 2009). As already noted above, it is known that at least miRNA and piRNA genes proliferated and diversified via gene duplication (for lncRNAs there is so far little evidence). But how did the original noncoding RNA genes arise? What are their ancestral precursors? Could they also have evolved *de novo* from previously nonfunctional genomic sequence, akin to the protein-coding genes described above? Recent work has started to provide some pertinent answers to these questions.

***Long noncoding RNA origination from scratch.*** A recent pioneering study dissected the origin and functional implications of a multi-exonic lncRNA in mice (Heinen et al. 2009). The gene expressing this RNA, *Poldi*, seems to have arisen through the transcriptional activation of a region containing preexisting cryptic splice sites in post-meiotic testis cells (spermatids) and was fixed by a selective sweep in *M. m. musculus* populations. Remarkably, knocking out *Poldi* led to reduced sperm motility and reduced testis weight, suggesting that *Poldi* contributed to enhanced fertility of the mice carrying it. Gene expression analyses indicate that the molecular basis of this phenotype is related to regulatory changes at the chromatin level induced by this new RNA gene, in line with the

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



notion that lncRNA often exert regulatory functions (Ponting et al. 2009). Given the pervasive transcription of the genome and the fact that useable proto-promoters (or promoters that can be co-opted from other genes) and cryptic splice sites abound in the genome (as also evidenced by the emergence of multi-exonic retrogenes; (Kaessmann et al. 2009); see above), *de novo* emergence of non-coding RNA (Fig. 4A) genes as exemplified by *Poldi* might turn out to be a rather frequent phenomenon. However, the regulatory, sequence and structural requirements for the functionality of long noncoding RNAs are so far poorly understood and hence the probability of such gene formation events is hard to predict.

***Protein-coding genes transformed into RNA genes.*** The origin of a classic lncRNA gene suggests an important alternative trajectory for the origin of new lncRNAs (Fig. 4B). The *Xist* gene, well-known for its crucial role in X chromosome dosage compensation in eutherian mammals (where it triggers transcriptional inactivation of one female X chromosome), emanated from the remnants of a former protein-coding gene (Duret et al. 2006). This metarmophosis involved the loss of protein-coding capacity of the precursor gene's exons and subsequent reuse of several of these exons and original promoter elements in the newly minted *Xist* RNA gene. But the origin of lncRNA genes from protein-coding antecedents is not confined to mammals. An intriguing example from *Drosophila* is the *spx* gene, which represents a fusion of an *ATP synthase* gene to functionally uncharacterized exons near the insertion site (Wang et al. 2002). Remarkably, the *spx* ancestor lost its coding capacity and evolved into an RNA gene with a function in male courtship behavior, a process that was shaped by positive selection. These cases illustrate that the formation of new lncRNA genes may directly draw from previous gene structure information and regulatory capacity. Given the constant generation of new protein-coding gene copies through gene duplication and the frequent (often associated) gene death processes during evolution, the origin of *Xist* and *spx* might exemplify a potentially common mechanism.

***Small RNAs.*** The birth of small RNAs also seems to have benefited from erstwhile protein-coding gene material. For example, two primate miRNA genes were shown to have arisen from retropseudogenes, a process that apparently profited from the fact that the pseudogenes provided sequences of the potential target genes (the retropseudogenes' parental genes) and regulatory elements (Devor 2006). Similarly, but at a larger scale, it was found that mammalian retropseudogenes seem to frequently encode small interfering RNAs that may play important roles in the regulation of their parental source genes in the germline (Tam et al. 2008; Watanabe et al. 2008).

## New genes from domesticated genomic parasites

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



Parasitic elements of the genome, such as transposons and endogenous retroviruses, have indirectly contributed to the functional evolution of genomes in many ways. For example, given that transposable elements are key mediators of segmental duplication (by stimulating various recombination events; Fig. 1A; (Marques-Bonet et al. 2009a)) and provide the core machinery underlying retroduplication (see above), they represent primary promoters of new gene birth. But, interestingly, genomic parasites have also more directly contributed to the evolution of new genes in their host genomes, as summarized in the following.

***Protein-coding genes from genome parasites.*** It has been known for quite some time that transposable elements have frequently been incorporated into genes as new exons, a process frequently associated with alternative splicing (Sorek 2007). However, the functional significance of these "exonization" events has remained elusive. More strikingly, a number of new genes that were, by and large, entirely derived from genome "parasites" and evolved beneficial functions for the host organism have been identified in recent years (Volff 2006). Examples for such "domesticated" parasites are the *syncytin* genes, which stem from envelope genes of endogenous retroviruses and originated independently in primates, rodents, and lagomorphs (Fig. 5; (Mi et al. 2000; Dupressoir et al. 2009; Heidmann et al. 2009)). Remarkably, in all of these mammalian lineages, the syncytin-encoded proteins were co-opted to mediate crucial functions in placentation. That is, they are essential for the development of the "syncytium", an exterior structure of the placenta that is essential for proper nutrient and waste exchange between mother and fetus. Thus, the eutherian placenta, a recent evolutionary innovation, appears to have provided a particularly fruitful ground for the emergence of new domesticated genes with beneficial functions, a view that is further supported by the observation that two retrotransposon-derived genes (*Peg10* and *Peg11*) have similarly adopted key functional roles in the murine placenta (Ono et al. 2006; Sekita et al. 2008).

However, other functional roles have been assigned to "tamed" genomic parasites as well. For instance, a recent study traced the birth of a new transcription factor gene (*ZBED6*) back to the domestication of a DNA transposon in the common ancestor of eutherians (Markljung et al. 2009). ZBED6 has evolved key regulatory roles in muscle growth, but, interestingly, may affect the expression of thousands of other genes that control fundamental biological processes and therefore could underlie the evolution of a completely new regulatory network in placental mammals.

***Noncoding RNAs from transposable elements.*** In addition to various other protein-coding genes that arose on the basis of transposable element sequences in diverse taxa (i.e., vertebrates, fruitflies, and plants; (Volff 2006)), several long and small RNA genes were shown to represent "reincarnated" retrotransposons. This process is exemplified by the origin of the brain cytoplasmic lncRNA genes (*BC1* and *BC200*). Although these genes evolved

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



independently from retrotransposons in rodents and anthropoid primates (Brosius 1999), they adapted to similar roles in translational regulation in the brain (Cao et al. 2006). While cases of lncRNAs that were derived from transposon ancestors are so far scarce, new small RNA genes seem to rather frequently have emerged from transposable elements. For example, retrotransposon conversions have given rise to dozens of known lineage-specific miRNAs in mammals (Smalheiser and Torvik 2005; Piriyapongsa et al. 2007). Finally, the germline-expressed piRNAs and endo-siRNAs should also be mentioned in this context, because they are frequently derived from the various lineage-specific transposable elements that they then control (Malone and Hannon 2009).

## Horizontal gene transfer

Horizontal gene transfer (HGT; also known as lateral gene transfer) is the process by which an organism incorporates genetic material from another organism without being a direct descendant of that organism. The importance of HGT in bacterial evolution is long established (Boucher et al. 2003). HGT has also been frequently documented in phagocytic and parasitic unicellular eukaryotes (Keeling and Palmer 2008). However, until recently, HGT involving animals and plants appeared to be confined to events associated with endosymbiosis (e.g., transfer of mitochondrial or plastid genes to the nuclear genome) or parasitism (e.g., transfer of genes from the intracellular Wolbachia bacteria to their *Drosophila* hosts (Hotopp et al. 2007)). It is thought that HGT is limited in animals because of a highly segregated and sheltered germline (Keeling and Palmer 2008). Interestingly, however, a recent study revealed that a species of rotifers (wheel animals) has acquired numerous genes from various other organisms (i.e., bacteria, fungi and plants), potentially associated with the extreme environmental stress (repeated desiccation) to which this organism is subjected (Gladyshev et al. 2008). However, although several acquired genes seem to have remained intact, the functional relevance of this curious case of HGT still needs to be established.

## The testis: a catalyst for the birth and evolution of new genes in animals?

Collectively, studies of new genes in animals have ascribed one specific organ an intriguing and potentially central role in the process of gene birth and evolution. Probably not fortuitously, already the first detailed investigations of recent gene origination in mammals (*Pgk-2*; (McCarrey and Thomas 1987)) and *Drosophila* (*jingwei*, (Long and Langley 1993)) revealed the newly formed genes to be specifically expressed in one tissue: the testis (the earlier individual examples are also reviewed in (Brosius 1999)). Global studies of retroduplication later showed an overall propensity of young retrogenes to be expressed in this organ in these species (Betran et al. 2002; Marques et al. 2005). Based on these observations, we suggested that the testis may represent a crucible for new gene evolution (Fig. 6), allowing novel genes to form and evolve, and potentially adopt functions in other (somatic) tissues with time (Marques et al. 2005; Vinckenbosch et al. 2006; Kaessmann et al. 2009).

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



Indeed, as may also have transpired from the various examples discussed in this review, an increasing body of literature highlights that young new genes of all kinds seem to have been preferentially endowed with testis-specific expression patterns and/or functional roles in this tissue during evolution. Thus, not only retrogenes but also young (partial) segmental duplicates, chimeric genes, as well as protein-coding and RNA genes that emerged *de novo* in mammals and fruitflies have often been found to show testis-specific or testis-biased transcription (e.g., refs (Paulding et al. 2003; She et al. 2004; Levine et al. 2006; Heinen et al. 2009); see also examples discussed above). Although these individual observations only indicate a strong general trend that needs to be verified on a large scale for the different types of newly shaped genes, this raises the question of why the testis might provide a common evolutionary conduit for the fixation and functional evolution of new genes.

Several factors likely contributed to the "out of the testis" emergence of new genes. It is well established that, at the genomic and molecular level, the testis constitutes the most rapidly evolving organ, owing to the intense selective pressures to which it is subjected and which are associated with sperm competition, sexual conflict, reproductive isolation, germline pathogens, and mutations causing segregation distortion in the male germline (Nielsen et al. 2005). Thus, the testis may represent an evolutionarily "greedy" tissue, highly receptive for the accommodation of evolutionary genomic innovations such as new genes.

But which factor allows for the (specific) transcription of so many new gene structures in this tissue in the first place, the prerequisite for regularly "feeding" the testis with functional new gene material during evolution? The answer to this question, at least in mammals, can potentially be sought in the peculiar properties of transcription in the meiotic and postmeiotic spermatogenic cells, termed spermatocytes and spermatids, respectively (Fig. 6). Thus, a number of molecular analyses of individual genes suggest that various specific histone variants and modifications might favor an open chromatin conformation in these cells (Kleene 2001; Sassone-Corsi 2002; Kimmins and Sassone-Corsi 2005). Together with widespread demethylation of CpG-enriched promoter elements and potentially elevated levels of core components of the transcriptional machinery (Kleene 2001), these factors may lead to a "promiscuous" or permissive state of chromatin in spermatocytes and spermatids, which might imply widespread transcription of nonfunctional or otherwise not transcribed genomic elements. We thus speculated that this specific chromatin state might have facilitated the initial, promiscuous transcription of newly arisen gene copies (that may often initially lack powerful regulatory elements) in the testis during their early evolution (Marques et al. 2005; Vinckenbosch et al. 2006; Kaessmann et al. 2009). A subset of these new gene candidates subsequently obtained beneficial functions in these germ cells and evolved into *bona fide* genes, a process that perhaps was further facilitated by the fact that efficient and specific expression in these germ cells may require only relatively simple (CpG-enriched) promoters (Kleene 2005), which may potentially arise *de novo* through

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



relatively few mutational steps. Natural selection then further refined the promoters of these new genes, which may ultimately also have led to expression and beneficial functions in other (somatic) tissues, although functions of many of these new genes may have remained restricted to the rapidly evolving testis.

The existence of a chromosome-wide promiscuous state of autosomal chromatin that would favor the transcriptional activity of new gene structures in intergenic regions remains to be validated. Also, as indicated above, future studies need to confirm the potential testis-bias for a larger number of new genes of the different classes. These studies should also systematically compare spatial expression patterns of young and old categories of new genes, to assess whether expression in testes really represents a regular and general catalyst for new gene origination by providing an entranceway for the evolution of new gene functions even in other tissues.

## Conclusions and future prospects

Although the origin of the first, primordial genes may ultimately be traced back to some precursors in the so called "RNA world" billions of years ago (Gilbert 1986), their origins remain enigmatic. By contrast, a rapidly increasing number of studies facilitated by the genomics era and based on extant genome sequences have revealed an astounding diversity of mechanisms underlying the birth of more recent genes. Almost any imaginable pathway towards new gene birth seems to have been documented by now, even those previously deemed highly unlikely or impossible. Thus, new genes have arisen from copies of old ones, protein and RNA genes were composed from scratch, protein-coding genes metamorphosed into RNA genes, parasitic genome sequences were domesticated and, finally, all of the resulting components also readily mixed to yield new chimeric genes with unprecedented functions. On top of that, several of these rather rare trajectories of new gene birth and evolution were demonstrated to have reoccurred independently in separate evolutionary lineages.

Together, these observations illustrate that even rare gene formation events can be driven to fixation during evolution, provided that the selective benefits are high enough. Indeed, new genes of various types were demonstrated to have imparted numerous favorable functional and phenotypic innovations. They have significantly impacted the evolution of cellular, physiological, morphological, behavioral, and reproductive phenotypic traits. Therefore, as had long been surmised (Ohno 1970), it is now beyond doubt that new genes have significantly contributed to organismal evolution.

***Future directions.*** So, what remains to be done? Overall, in spite of the tremendous progress in the field, it should be conceded that the repertoire of *bona fide* new functional genes in most organisms is so far, overall, rather poorly characterized at the functional level. This is especially true with respect to more recently emerged cases,

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



which can be considered to be particularly interesting, given that they may have contributed to evolutionary change and phenotypic adaptation in more recently diverged evolutionary lineages. The present limitation is due to challenges in pinpointing relevant functional cases among the vast number of gene copies and other genomic constituents (e.g., transposable elements) that may represent concealed new genes, as well as the difficulty in unraveling their functions on a decent scale.

Thus, future work should aim to identify a larger number of new gene candidates, an endeavor that will be greatly facilitated by the recent availability of revolutionary high throughput sequencing technologies, which allow the generation of genome and transcriptome data for numerous organisms at an unprecedented pace and scale (Wang et al. 2008; Metzker 2010). Specifically, these new data will facilitate the evolutionary analysis of genomes and transcriptomes from many different organisms and hence greatly accelerate the discovery of new gene and transcript structures (e.g., of otherwise hard-to-detect lncRNA loci) as well as the selective forces that may have shaped these genes and their transcriptional activities. Thus, it will be possible to more precisely assess the contribution of the different mechanisms underlying the formation of novel gene structures (many of which are so far of rather anecdotal nature) and the resulting proportions of different types of new genes. These new data will likely also lead to the discovery of unanticipated novel modes of new gene origination. Generally, given recent developments and observations in the field, the identification and characterization of genes that emerged *de novo* as well as the detection and analysis of new noncoding RNA genes may represent particularly exciting and fruitful areas of future investigation.

However, future efforts should more often strive to go beyond the mere description of new gene structures and their evolutionary and selective signatures. Although challenging, newly identified novel genes should be subjected to in-depth characterizations of their functional evolution, using evolutionary analysis combined with large- and small-scale genomics/transcriptomics, molecular, cellular and *in vivo* experiments. Such studies may elucidate, at a larger scale, the molecular changes associated with the evolution of novel functions that emerged from the different types of new genes. For instance, they could investigate in more detail the relative roles of gene expression change, protein/RNA sequence divergence and subcellular relocalizations of the encoded gene products in this process. They should also help to clarify the generality of previous observations and hypotheses concerning the functional evolution of new genes, such as the evolutionary role of the testis in the emergence of new genes and their functions. Finally, such detailed functional investigations will establish the biological relevance for a greater number of newly minted genes. Future work will therefore ultimately reveal the contribution of the process of new gene birth to the evolution of adaptive evolutionary novelties relative to that provided by evolutionary alterations and fine-tuning of long-existing ancestral genes.

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



## Acknowledgments

I apologize to colleagues whose work could not be discussed or cited owing to space constraints and/or the focus of this review. I thank the three anonymous reviewers for helpful suggestions. I thank the members of my lab for helpful discussions and David Brawand and Nihal Okaya for advice regarding the generation of the figures. This work was supported by grants from the Swiss National Science Foundation and the European Research Council.

## References

Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, Novik A, and Sorek R. 2006. Transcription-mediated gene fusion in the human genome. *Genome Res* **16**(1): 30-36.

Assis R and Kondrashov AS. 2009. Rapid repetitive element-mediated expansion of piRNA clusters in mammalian evolution. *Proc Natl Acad Sci U S A* **106**(17): 7079-7082.

Babushok DV, Ohshima K, Ostertag EM, Chen X, Wang Y, Mandal PK, Okada N, Abrams CS, and Kazazian HH, Jr. 2007. A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids. *Genome Res* **17**(8): 1129-1138.

Bai Y, Casola C, Feschotte C, and Betran E. 2007. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila. *Genome Biol* **8**(1): R11.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, and Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **297**(5583): 1003-1007.

Betran E and Long M. 2003. Dntf-2r, a young Drosophila retroposed gene with specific male expression under positive Darwinian selection. *Genetics* **164**(3): 977-988.

Betran E, Thornton K, and Long M. 2002. Retroposed new genes out of the x in Drosophila. *Genome Res* **12**(12): 1854-1859.

Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau ME, Nesbo CL, Case RJ, and Doolittle WF. 2003. Lateral gene transfer and the origins of prokaryotic groups. *Annu Rev Genet* **37**: 283-328.

Bradley J, Baltus A, Skaletsky H, Royce-Tolland M, Dewar K, and Page DC. 2004. An X-to-autosome retrogene is required for spermatogenesis in mice. *Nat Genet* **36**(8): 872-876.

Brennan G, Kozyrev Y, and Hu SL. 2008. TRIMCyp expression in Old World primates Macaca nemestrina and Macaca fascicularis. *Proc Natl Acad Sci U S A* **105**(9): 3569-3574.

Brosius J. 1991. Retroposons--seeds of evolution. *Science* **251**(4995): 753.

Brosius J. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* **238**(1): 115-134.

Burki F and Kaessmann H. 2004. Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nature Genet* **36**(10): 1061-1063.

Cai J, Zhao R, Jiang H, and Wang W. 2008. De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. *Genetics* **179**(1): 487-496.

Cao X, Yeo G, Muotri AR, Kuwabara T, and Gage FH. 2006. Noncoding RNAs in the mammalian central nervous system. *Annu Rev Neurosci* **29**: 77-103.

Carthew RW and Sontheimer EJ. 2009. Origins and Mechanisms of miRNAs and siRNAs. *Cell* **136**(4): 642-655.

Ciccarelli FD, von Mering C, Suyama M, Harrington ED, Izaurralde E, and Bork P. 2005. Complex genomic rearrangements lead to novel primate gene function. *Genome Res* **15**(3): 343-351.

Conant GC and Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet* **9**(12): 938-950.

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



Cordaux R and Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**(10): 691-703.

Demuth JP and Hahn MW. 2009. The life and death of gene families. *Bioessays* **31**(1): 29-39.

Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T, Manzano C, Chrast J et al. 2007. Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res* **17**(6): 746-759.

Devor EJ. 2006. Primate microRNAs miR-220 and miR-492 lie within processed pseudogenes. *J Hered* **97**(2): 186-190.

Devor EJ and Samollow PB. 2008. In vitro and in silico annotation of conserved and nonconserved microRNAs in the genome of the marsupial Monodelphis domestica. *J Hered* **99**(1): 66-72.

Dupressoir A, Vernochet C, Bawa O, Harper F, Pierron G, Opolon P, and Heidmann T. 2009. Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. *Proc Natl Acad Sci U S A* **106**(29): 12127-12132.

Duret L, Chureau C, Samain S, Weissenbach J, and Avner P. 2006. The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* **312**(5780): 1653-1655.

Emerson JJ, Kaessmann H, Betran E, and Long M. 2004. Extensive gene traffic on the mammalian X chromosome. *Science* **303**(5657): 537-540.

Emes RD, Goodstadt L, Winter EE, and Ponting CP. 2003. Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum Mol Genet* **12**(7): 701-709.

Esnault C, Maestre J, and Heidmann T. 2000. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet* **24**(4): 363-367.

Fablet M, Bueno M, Potrzebowski L, and Kaessmann H. 2009. Evolutionary origin and functions of retrogene introns. *Mol Biol Evol* **26**(9): 2147-2156.

Feng Q, Moran JV, Kazazian HH, Jr., and Boeke JD. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**(5): 905-916.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, and Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**(4): 1531-1545.

Gherman A, Chen PE, Teslovich TM, Stankiewicz P, Withers M, Kashuk CS, Chakravarti A, Lupski JR, Cutler DJ, and Katsanis N. 2007. Population bottlenecks as a potential major shaping force of human genome architecture. *PLoS Genet* **3**(7): e119.

Gilbert W. 1986. Origin of life: The RNA world. *Nature* **319**: 618.

Gladyshev EA, Meselson M, and Arkhipova IR. 2008. Massive horizontal gene transfer in bdelloid rotifers. *Science* **320**(5880): 1210-1213.

Guo X, Su B, Zhou Z, and Sha J. 2009. Rapid evolution of mammalian X-linked testis microRNAs. *BMC Genomics* **10**: 97.

Haldane JBS. 1933. The part played by recurrent mutation in evolution. *Amer Natur,* **67**: 5-9.

Heidmann O, Vernochet C, Dupressoir A, and Heidmann T. 2009. Identification of an endogenous retroviral envelope gene with fusogenic activity and placenta-specific expression in the rabbit: a new "syncytin" in a third order of mammals. *Retrovirology* **6**: 107.

Heinen TJ, Staubach F, Haming D, and Tautz D. 2009. Emergence of a new gene from an intergenic region. *Curr Biol* **19**(18): 1527-1531.

Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, Hofacker IL, and Stadler PF. 2006. The expansion of the metazoan microRNA repertoire. *BMC Genomics* **7**: 25.

Hillier LW Miller W Birney E Warren W Hardison RC Ponting CP Bork P Burt DW Groenen MA Delany ME et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**(7018): 695-716.

Hotopp JC, Clark ME, Oliveira DC, Foster JM, Fischer P, Torres MC, Giebel JD, Kumar N, Ishmael N, Wang S et al. 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* **317**(5845): 1753-1756.

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



Innan H and Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet* **11**(2): 97-108.

Jacob F. 1977. Evolution and tinkering. *Science* **196**(4295): 1161-1166.

Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, and Eichler EE. 2001. Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**(6855): 514-519.

Kaessmann H, Vinckenbosch N, and Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* **10**: 19-31.

Keeling PJ and Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet* **9**(8): 605-618.

Keightley PD, Lercher MJ, and Eyre-Walker A. 2005. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol* **3**(2): e42.

Kimmins S and Sassone-Corsi P. 2005. Chromatin remodelling and epigenetic features of germ cells. *Nature* **434**(7033): 583-589.

Kleene KC. 2001. A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. *Mech Dev* **106**(1-2): 3-23.

Kleene KC. 2005. Sexual selection, genetic conflict, selfish genes, and the atypical patterns of gene expression in spermatogenic cells. *Dev Biol* **277**(1): 16-26.

Knowles DG and McLysaght A. 2009. Recent de novo origin of human protein-coding genes. *Genome Res* **19**(10): 1752-1759.

Levine MT, Jones CD, Kern AD, Lindfors HA, and Begun DJ. 2006. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proc Natl Acad Sci U S A* **103**(26): 9935-9939.

Li WH. 1997. *Molecular Evolution*. Sinauer Associates, Sunderland MA.

Liu SL, Zhuang Y, Zhang P, and Adams KL. 2009. Comparative analysis of structural diversity and sequence evolution in plant mitochondrial genes transferred to the nucleus. *Mol Biol Evol* **26**(4): 875-891.

Long M, Betran E, Thornton K, and Wang W. 2003. The origin of new genes: glimpses from the young and old. *Nat Rev Genet* **4**(11): 865-875.

Long M and Langley CH. 1993. Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. *Science* **260**(5104): 91-95.

Lynch M. 2007. *The origins of genome architecture.* Sinauer Associates., Sunderland, USA.

Lynch M and Conery JS. 2003. The evolutionary demography of duplicate genes. *J Struct Funct Genomics* **3**(1-4): 35-44.

Malone CD and Hannon GJ. 2009. Small RNAs as guardians of the genome. *Cell* **136**(4): 656-668.

Markljung E, Jiang L, Jaffe JD, Mikkelsen TS, Wallerman O, Larhammar M, Zhang X, Wang L, Saenz-Vash V, Gnirke A et al. 2009. ZBED6, a novel transcription factor derived from a domesticated DNA transposon regulates IGF2 expression and muscle growth. *PLoS Biol* **7**(12): e1000256.

Marques A, Dupanloup I, Vinckenbosch N, Reymond A, and Kaessmann H. 2005. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol* **3**(11): e357.

Marques AC, Vinckenbosch N, Brawand D, and Kaessmann H. 2008. Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. *Genome Biol* **9**(3): R54.

Marques-Bonet T and Eichler EE. 2009. The Evolution of Human Segmental Duplications and the Core Duplicon Hypothesis. *Cold Spring Harb Symp Quant Biol*.

Marques-Bonet T, Girirajan S, and Eichler EE. 2009a. The origins and impact of primate segmental duplications. *Trends Genet* **25**(10): 443-454.

Marques-Bonet T, Ryder OA, and Eichler EE. 2009b. Sequencing primate genomes: what have we learned? *Annu Rev Genomics Hum Genet* **10**: 355-386.

Mathias SL, Scott AF, Kazazian HH, Jr., Boeke JD, and Gabriel A. 1991. Reverse transcriptase encoded by a human transposable element. *Science* **254**(5039): 1808-1810.

McCarrey JR and Thomas K. 1987. Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. *Nature* **326**(6112): 501-505.

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



Mercer TR, Dinger ME, and Mattick JS. 2009. Long non-coding RNAs: insights into functions. *Nat Rev Genet* **10**(3): 155-159.

Metzker ML. 2010. Sequencing technologies - the next generation. *Nat Rev Genet* **11**(1): 31-46.

Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard P, Howes S et al. 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403**(6771): 785-789.

Mighell AJ, Smith NR, Robinson PA, and Markham AF. 2000. Vertebrate pseudogenes. *FEBS Lett* **468**(2-3): 109-114.

Moran JV, DeBerardinis RJ, and Kazazian HH, Jr. 1999. Exon shuffling by L1 retrotransposition. *Science* **283**(5407): 1530-1534.

Muller HJ. 1935. The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere. *Genetics* **17**: 237-252.

Murchison EP, Kheradpour P, Sachidanandam R, Smith C, Hodges E, Xuan Z, Kellis M, Grutzner F, Stark A, and Hannon GJ. 2008. Conservation of small RNA pathways in platypus. *Genome Res* **18**(6): 995-1004.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ et al. 2005. A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees. *PLoS Biol* **3**(6): e170.

Nugent JM and Palmer JD. 1991. RNA-mediated transfer of the gene coxII from the mitochondrion to the nucleus during flowering plant evolution. *Cell* **66**(3): 473-481.

Ohno S. 1970. *Evolution by Gene Duplication*. Springer Verlag, Berlin.

Ohno S. 1972. So much "junk" DNA in our genome. In *Evolution of Genetic Systems*, Vol 23. Brookhaven Symp. Biol.

Okamura K and Nakai K. 2008. Retrotransposition as a source of new promoters. *Mol Biol Evol* **25**(6): 1231-1238.

Ono R, Nakamura K, Inoue K, Naruse M, Usami T, Wakisaka-Saito N, Hino T, Suzuki-Migishima R, Ogonuki N, Miki H et al. 2006. Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. *Nat Genet* **38**(1): 101-106.

Parker HG, VonHoldt BM, Quignon P, Margulies EH, Shao S, Mosher DS, Spady TC, and Elkahloun A. 2009. An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science* **0**: 0-0.

Parra G, Reymond A, Dabbouseh N, Dermitzakis ET, Castelo R, Thomson TM, Antonarakis SE, and Guigo R. 2006. Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res* **16**(1): 37-44.

Patthy L. 1999. *Protein Evolution*. Blackwell Science, Oxford.

Paulding CA, Ruvolo M, and Haber DA. 2003. The Tre2 (USP6) oncogene is a hominoid-specific gene. *Proc Natl Acad Sci U S A* **100**(5): 2507-2511.

Piriyapongsa J, Marino-Ramirez L, and Jordan IK. 2007. Origin and evolution of human microRNAs from transposable elements. *Genetics* **176**(2): 1323-1337.

Ponting CP, Oliver PL, and Reik W. 2009. Evolution and functions of long noncoding RNAs. *Cell* **136**(4): 629-641.

Potrzebowski L, Vinckenbosch N, and Kaessmann H. 2010. The emergence of new genes on the young therian X. *Trends Genet* **26**(1): 1-4.

Potrzebowski L, Vinckenbosch N, Marques AC, Chalmel F, Jegou B, and Kaessmann H. 2008. Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol* **6**(4): e80.

Pradet-Balade B, Medema JP, Lopez-Fraga M, Lozano JC, Kolfschoten GM, Picard A, Martinez AC, Garcia-Sanz JA, and Hahne M. 2002. An endogenous hybrid mRNA encodes TWE-PRIL, a functional cell surface TWEAK-APRIL fusion protein. *EMBO J* **21**(21): 5711-5720.

Rohozinski J and Bishop CE. 2004. The mouse juvenile spermatogonial depletion (jsd) phenotype is due to a mutation in the X-derived retrogene, mUtp14b. *Proc Natl Acad Sci U S A* **101**(32): 11695-11700.

Romero D and Palacios R. 1997. Gene amplification and genomic plasticity in prokaryotes. *Annu Rev Genet* **31**: 91-111.

Rosso L, Marques AC, Reichert AS, and Kaessmann H. 2008a. Mitochondrial targeting adaptation of the hominoid-specific glutamate dehydrogenase driven by positive Darwinian selection. *PLoS Genet* **4**(8): e1000150.

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



Rosso L, Marques AC, Weier M, Lambert N, Lambot M-A, Vanderhaeghen P, and Kaessmann H. 2008b. Birth and Rapid Subcellular Adaptation of a Hominoid-Specific CDC14 Protein. *PLoS Biol* **6**(6): e140.

Sassone-Corsi P. 2002. Unique chromatin remodeling and transcriptional regulation in spermatogenesis. *Science* **296**(5576): 2176-2178.

Sayah DM, Sokolskaja E, Berthoux L, and Luban J. 2004. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature* **430**(6999): 569-573.

Sekita Y, Wagatsuma H, Nakamura K, Ono R, Kagami M, Wakisaka N, Hino T, Suzuki-Migishima R, Kohda T, Ogura A et al. 2008. Role of retrotransposon-derived imprinted gene, Rtl1, in the feto-maternal interface of mouse placenta. *Nat Genet* **40**(2): 243-248.

She X, Cheng Z, Zollner S, Church DM, and Eichler EE. 2008. Mouse segmental duplication and copy number variation. *Nat Genet* **40**(7): 909-914.

She X, Horvath JE, Jiang Z, Liu G, Furey TS, Christ L, Clark R, Graves T, Gulden CL, Alkan C et al. 2004. The structure and evolution of centromeric transition regions within the human genome. *Nature* **430**(7002): 857-864.

Smalheiser NR and Torvik VI. 2005. Mammalian microRNAs derived from genomic repeats. *Trends Genet* **21**(6): 322-326.

Sorek R. 2007. The birth of new exons: mechanisms and evolutionary consequences. *RNA* **13**(10): 1603-1608.

Stern DL and Orgogozo V. 2009. Is genetic evolution predictable? *Science* **323**(5915): 746-751.

Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM et al. 2008. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**(7194): 534-538.

Thomson TM, Lozano JJ, Loukili N, Carrio R, Serras F, Cormand B, Valeri M, Diaz VM, Abril J, Burset M et al. 2000. Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene. *Genome Res* **10**(11): 1743-1756.

Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, and Alba MM. 2009. Origin of primate orphan genes: a comparative genomics approach. *Mol Biol Evol* **26**(3): 603-612.

Turner JM. 2007. Meiotic sex chromosome inactivation. *Development* **134**(10): 1823-1831.

Van de Peer Y, Maere S, and Meyer A. 2009. The evolutionary significance of ancient genome duplications. *Nat Rev Genet* **10**(10): 725-732.

Vibranovski MD, Lopes HF, Karr TL, and Long M. 2009. Stage-specific expression profiling of Drosophila spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic relocation of testis-expressed genes. *PLoS Genet* **5**(11): e1000731.

Vinckenbosch N, Dupanloup I, and Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci U S A* **103**(9): 3220-3225.

Volff JN. 2006. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays* **28**(9): 913-922.

Wang W, Brunet FG, Nevo E, and Long M. 2002. Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **99**(7): 4448-4453.

Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S et al. 2006. High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell* **18**(8): 1791-1802.

Wang Z, Gerstein M, and Snyder M. 2008. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**: 57-63.

Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T et al. 2008. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453**(7194): 539-543.

Xing J, Wang H, Belancio VP, Cordaux R, Deininger PL, and Batzer MA. 2006. Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proc Natl Acad Sci U S A* **103**(47): 17608-17613.

Zaiss DM and Kloetzel PM. 1999. 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**(5): 829-835.

Zhang J. 2003. Evolution by gene duplication. *Trends Ecol Evol* **18**: 292-298.

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



Zhang J. 2006. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nature Genet* **38**(7): 819-823.

Zhang J, Dean AM, Brunet F, and Long M. 2004. Evolving protein functional diversity in new genes of Drosophila. *Proc Natl Acad Sci U S A* **101**(46): 16246-16250.

Zhang J, Zhang YP, and Rosenberg HF. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nature Genet* **30**(4): 411-415.

Zhang R, Peng Y, Wang W, and Su B. 2007. Rapid evolution of an X-linked microRNA cluster in primates. *Genome Res* **17**(5): 612-617.

Zhang YW, Liu S, Zhang X, Li WB, Chen Y, Huang X, Sun L, Luo W, Netzer WJ, Threadgill R et al. 2009. A functional mouse retroposed gene Rps23r1 reduces Alzheimer's beta-amyloid levels and tau phosphorylation. *Neuron* **64**(3): 328-340.

Zhou Q and Wang W. 2008. On the origin and evolution of new genes--a genomic and experimental perspective. *J Genet Genomics* **35**(11): 639-648.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, and Wang W. 2008. On the origin of new genes in Drosophila. *Genome Res* **18**(9): 1446-1455.

Zhu Z, Zhang Y, and Long M. 2009. Extensive structural renovation of retrogenes in the evolution of the Populus genome. *Plant Physiol* **151**(4): 1943-1951.

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



## Figure legends

**Figure 1.** Origin of new gene copies through gene duplication. (*A*) DNA-based duplication. A common type of segmental duplication – tandem duplication – is shown. It may occur via unequal crossing-over that is mediated by transposable elements (light green). There are different fates of the resulting duplicate genes. For example, one of the duplicates may acquire new functions by evolving new expression patterns and/or novel biochemical protein or RNA functions (see main text for details). Exons are shown as golden or blue boxes (black connecting lines indicate exon splicing) and transcriptional start sites (TSSs) are red right-angled arrows. Nonexonic chromatin is depicted as grey tubes. (*B*) RNA-based duplication (termed retroposition or retroduplication). New retroposed gene copies may arise through the reverse transcription of messenger RNAs (mRNAs) from parental source genes. Functional retrogenes with new functional properties may evolve from these copies after acquisition or evolution of promoters in their 5' flanking regions that may drive their transcription (TSS is shown as transparent red angled arrow and additionally transcribed flanking sequence at the insertion site as transparent red box).

**Figure 2.** Origin of new chimeric gene or transcript structures. (*A*) DNA-based (genomic) gene fusion. Partial duplication (and hence fission) of ancestral source genes precedes juxtaposition of partial duplicates and subsequent fusion (presumably mediated by the evolution of novel splicing signals and/or transcription termination/polyadenylation sites). (*B*) Transcription-mediated gene fusion. Novel transcript structures may arise from intergenic splicing after evolution of novel splicing signals and transcriptional readthrough from the upstream gene. New chimeric mRNAs may sometimes be reversed transcribed to yield new chimeric retrogenes (see also Figure 1). Exons are shown as large boxes of different colors and transcriptional start sites (TSSs) are red right-angled arrows. Black connecting lines indicate constitutive splicing. Dotted lines indicate splicing of ancestral gene structures, whereas intergenic splicing that results in new chimeric transcripts is indicated by green lines.

**Figure 3.** Origin of protein-coding genes from scratch. New coding regions may emerge *de novo* from noncoding genomic sequences. First, proto-open reading frames (proto-ORFs; thin blue bars) acquire mutations (point substitutions, insertions/deletions; shown as yellow stars) that remove, bit by bit, frame-disrupting nucleotides (red wedges). Transcriptional activation of ORFs (through acquisition of promoters located in the 5' flanking region) encoding proteins with potentially useful functions may allow for the evolution of novel protein-coding genes (functional exon is shown as large blue box, TSS as right-angled arrow, and untranslated 5' sequence as transparent red box). Note that the transcriptional activation step may, alternatively, also precede the formation of complete functionally relevant ORFs.

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press



**Figure 4.** Evolutionary origins of long noncoding RNA genes. (*A*) *De novo* emergence. In this scenario, previously nonfunctional genomic sequence becomes transcribed (thin red box) through the acquisition/activation of a proto-promoter sequence (right-angled arrows). The transcriptional activation may be followed or preceded by the evolution of (proto-) splice sites (light blue stars). Together, these events allow for the formation of potentially functional and selectively beneficial multi-exonic noncoding RNA genes (exons are shown as large red boxes, splicing is indicated by thin black lines, TSSs are depicted as red right-angled arrows). (*B*) Origin of noncoding RNA gene from ancestral protein-coding gene. In this process, the original (functionally redundant) protein-coding gene loses its function and becomes a pseudogene. After or during loss of protein function and coding exon decay, a new functional noncoding RNA gene may arise, a process that may draw from regulatory elements and other sequences (splicing signals, exon sequences, polyadenylation sequences etc.) from the ancestral protein-coding gene. Protein-coding exons are shown as blue and RNA exons as red boxes (pseudogenized exons are shown as transparent boxes). Thin black lines indicate splicing, whereas lost ancestral splicing capacity is indicated by dotted lines. TSSs are shown as right-angled arrows.

**Figure 5.** New genes from domesticated genome parasites. The example shown illustrates the origin of a new placenta gene from an endogenous retrovirus sequence (the scenario illustrates the origin of one (Heidmann et al. 2009) of the several syncytin genes that evolved important placenta functions in mammals; see main text for details). The domestication event involved the decay of two of the human endogenous retrovirus ORFs (*gag* and *pol*; loss of function/decay is indicated by an empty box) and the selective preservation of the ORF encoding the virus envelope protein (ORFs depicted as golden boxes). The newly formed *syncytin* gene (transcript structure indicated by thin black line) became transcribed from the retrovirus' long terminal repeat (LTR; in green) promoter (TSS shown as right-angled red arrow) and evolved a placenta-specific expression pattern and (fusogenic) function (Heidmann et al. 2009).

**Figure 6.** The "out of the testis" hypothesis for the emergence of new genes. This hypothesis suggests that the transcription of new gene copies/structures (green boxes) is facilitated in certain testis germ cells – meiotic spermatocytes and post-meiotic round spermatids (which are found in the seminiferous tubules, where spermatogenesis takes place) – because of the potentially overall permissive chromatin state and overexpression of key components of the transcriptional machinery in these cells. The transcriptionally active chromatin state in spermatocytes and spermatids is thought to be a result of a potentially widespread demethylation of CpG dinucleotide-enriched promoter sequences and modifications (acetylation and methylation) of histones (blue ovals), which facilitate access of the transcriptional machinery (red ovals). Once transcribed, new functional genes (transcripts shown as green wavy lines) with beneficial products may be selectively preserved and evolve more

Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press

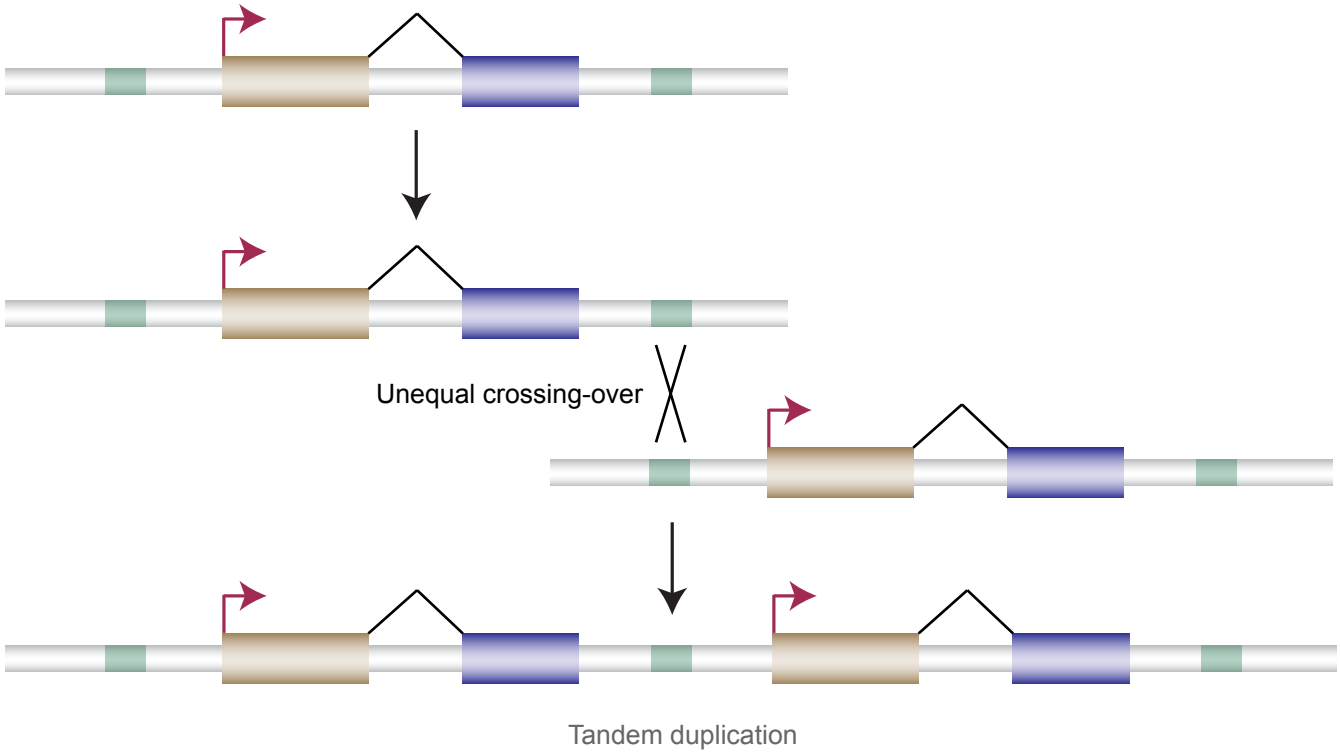


efficient promoters (a process that might be facilitated by the fact that spermatocyte/spermatid-specific expression require only relatively simple promoters). Eventually, such new genes may also evolve more diverse expression patterns and thus also obtain functions in other (somatic) tissues.

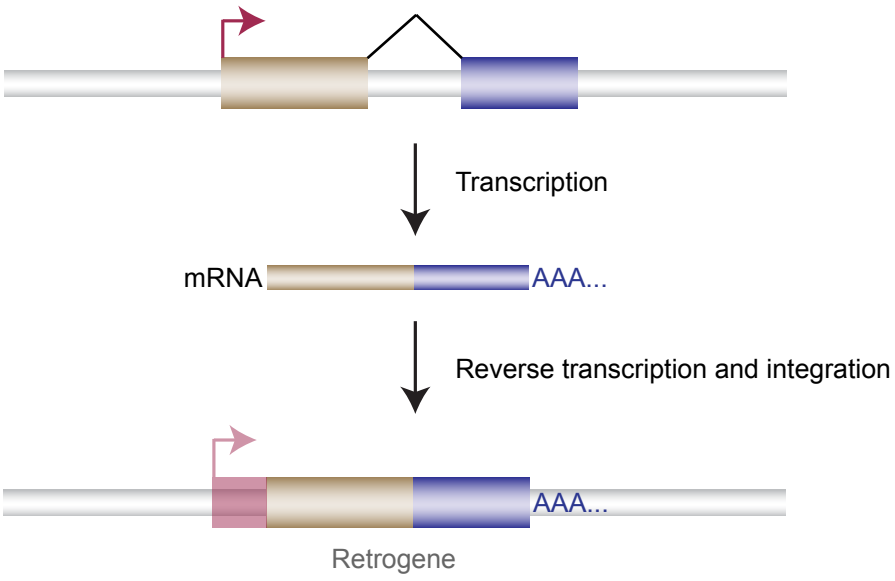
Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 1

A

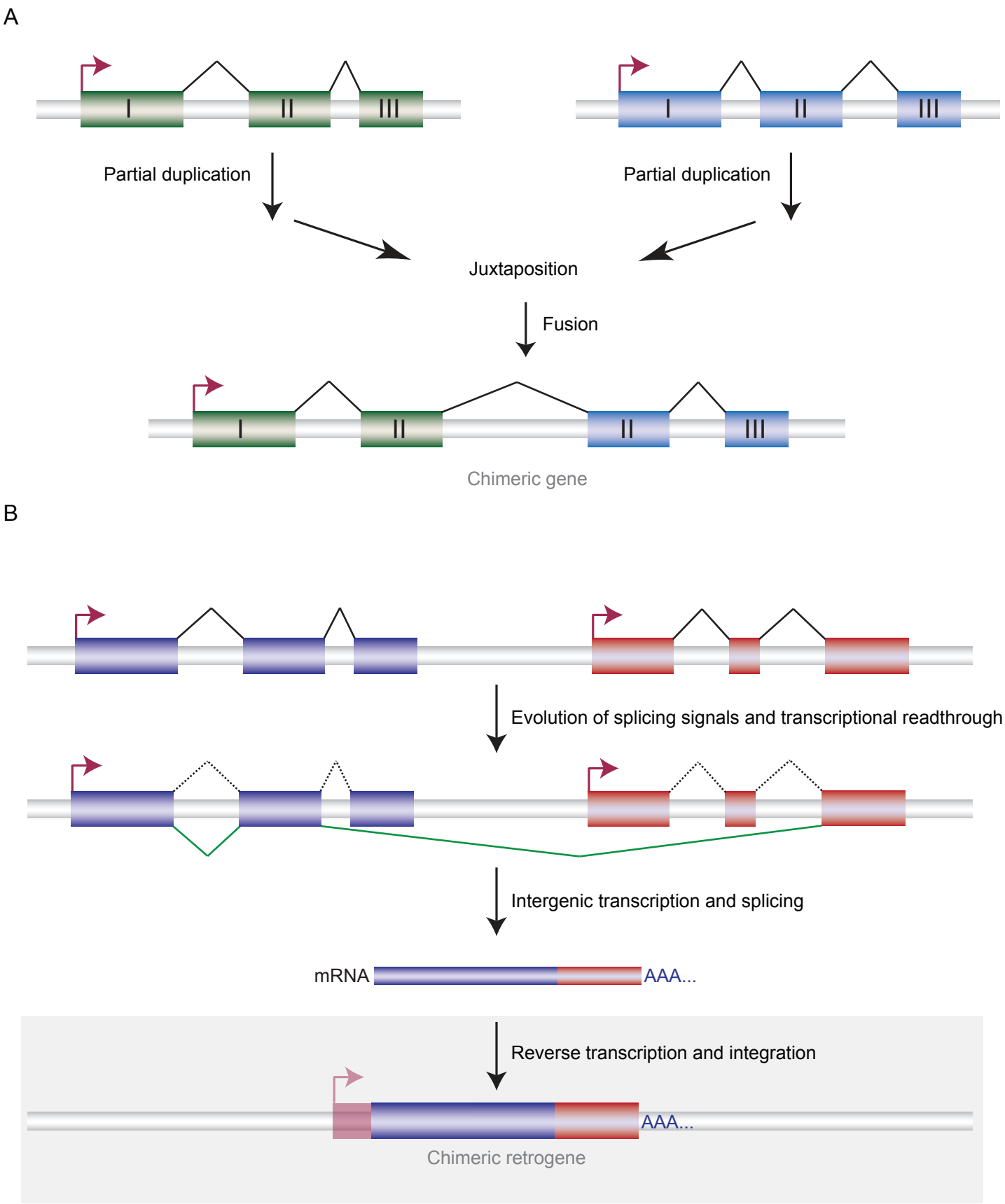



B



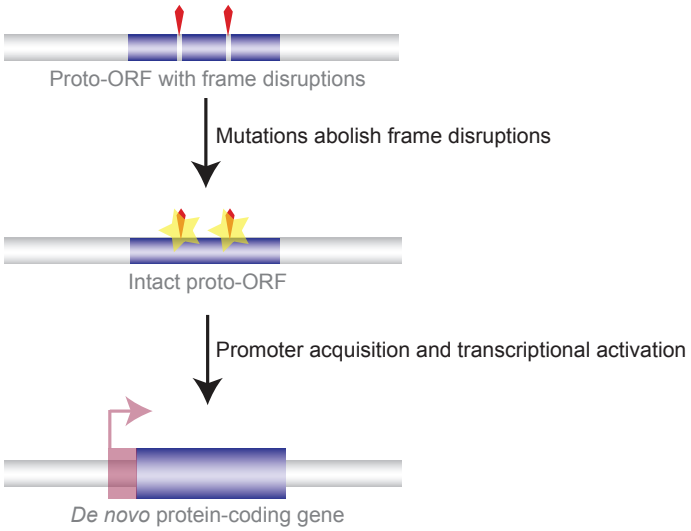
Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 2



Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press

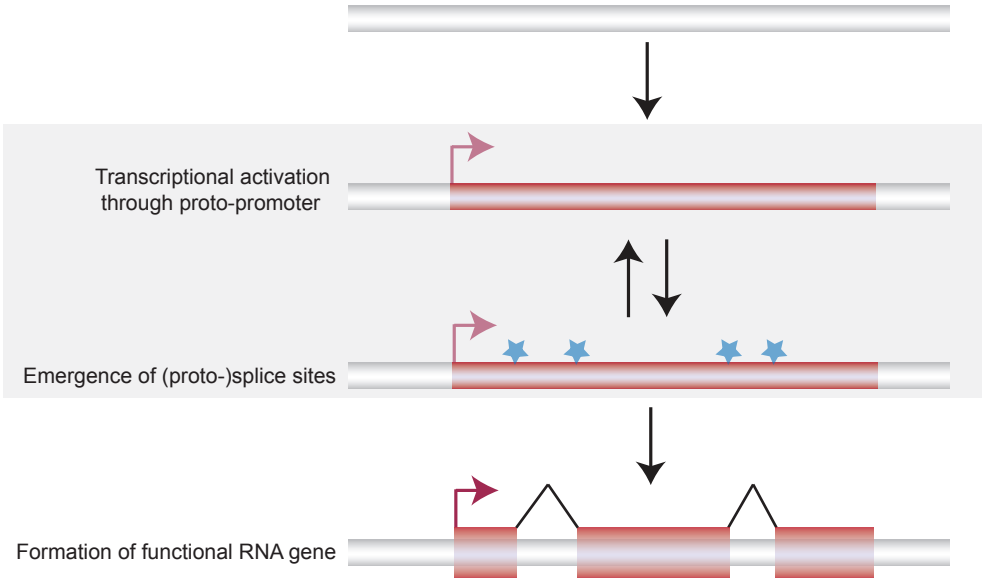
Figure 3



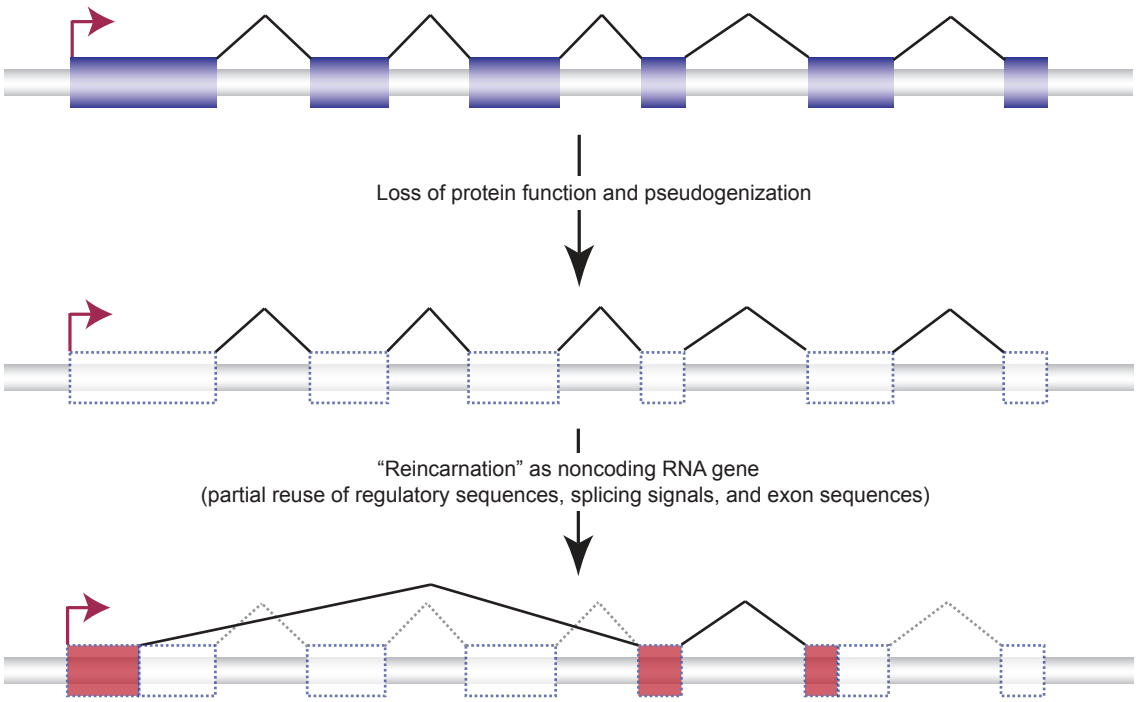
Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 4

A

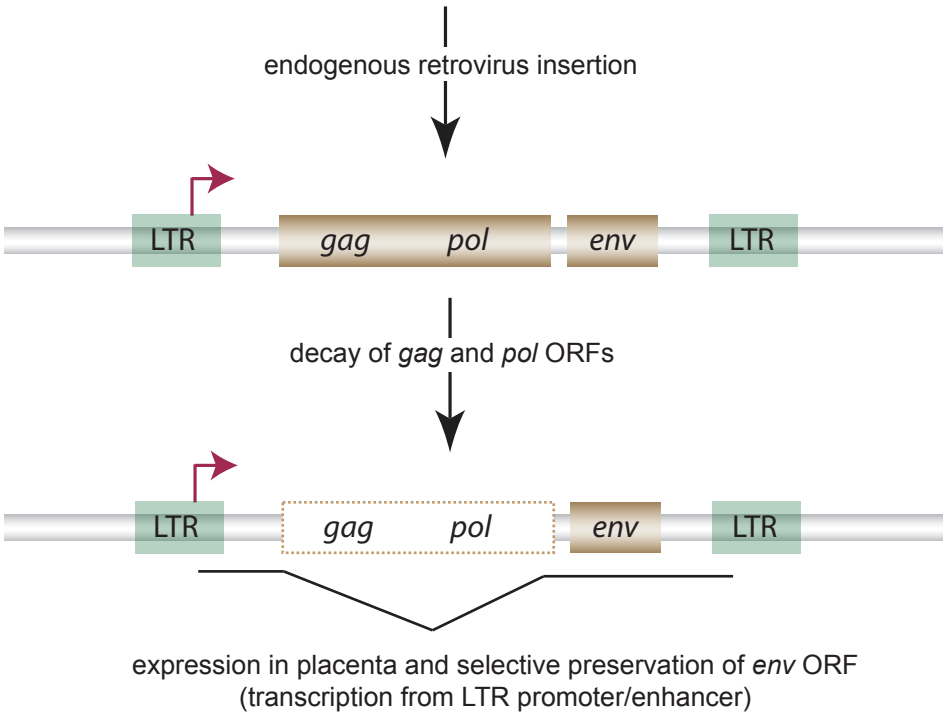



B



Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 5



Downloaded from genome.cshlp.org on June 14, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 6