Origins of New Genes and Evolution of Their Novel Functions

Yun Ding,[1,2] Qi Zhou,[3] and Wen Wang[1]

[1]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China; email: wwang@mail.kiz.ac.cn

[2]Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, Virginia 20147; email: dingy@janelia.hhmi.org

[3]Department of Integrative Biology, University of California, Berkeley, California 94709; email: zhouqi@berkeley.edu

Annu. Rev. Ecol. Evol. Syst. 2012. 43:345–63

First published online as a Review in Advance on September 4, 2012

The *Annual Review of Ecology, Evolution, and Systematics* is online at ecolsys.annualreviews.org

This article's doi: 10.1146/annurev-ecolsys-110411-160513

1543-592X/12/1201-0345$20.00

Keywords

origin of new genes, evolution of novel functions, de novo origination

Abstract

The origination of novel genes is an important process during the evolution of organisms because it provides critical sources for evolutionary innovation. Addressing how novel genes emerged and acquired novel and adaptive functions is of fundamental importance. Here we summarize the newest advances in our understanding of the molecular mechanisms and genome-wide patterns of new gene origination and new gene functions. We pay special attention to the origins of noncoding RNA genes and de novo genes, whose processes had been previously overlooked but are gaining increasingly visible importance. We then introduce recent findings that have opened a path to the study of the evolution of novel functions and pathways via novel genes. We also discuss the important issues and potential developments in the field.

1. INTRODUCTION

The origination of novel functional genes is one of the principal processes contributing to evolutionary innovation. Its significance was first recognized in the 1930s by Haldane (1935) and Muller (1933) and then formulated in Ohno's (1970) seminal monograph, *Evolution by Gene Duplication*. Ohno emphasized gene duplication as the most important mechanism producing new genes. Although this view might have been refined by recent studies (Levine et al. 2006, Zhou et al. 2008, Knowles & McLysaght 2009, Tautz & Domazet-Loso 2011, Wu et al. 2011), his idea that gene duplication provides functionally redundant new copies as raw material for natural selection has set the tone for subsequent theoretical and experimental studies. Under this hypothesis, new gene copies survive in the host genome in three main ways: preservation as a functionally redundant copy (Clark 1994), transformation into a complementary copy (subfunctionalization) (Lynch & Force 2000), or evolution into novel functions (neofunctionalization) (Walsh 1995). These models, especially the latter two, have been extensively documented in case studies of new genes within diverse organisms.

However, these models have been developed only for new genes that have another gene as its ancestor. Not until the recent development of whole-genome scans has another important source of new genes been recognized: de novo genes originating from noncoding ancestors (Levine et al. 2006, Zhou et al. 2008, Knowles & McLysaght 2009, Wu et al. 2011). Besides the exciting discovery of a novel mechanism, a more comprehensive picture of the origin and evolution of new genes is emerging in the genomic era (Zhou & Wang 2008, Kaessmann 2010). The abundant genomic resources now available allow investigators to identify and reconstruct an unbiased history of young genes (usually <10 million years) whose sequence changes can be clearly traced back to their parental sequences. The time at which a species diverges sets the upper limit for the ages of lineage-specific young genes. As such, they can be grouped into different ages within a phylogenetic context, which allows for a survey of the dynamic changes that new genes undergo over time. For example, two recent studies using 12 *Drosophila* genomes have found the patterns of origination mechanisms and gene expressions differ among new *Drosophila* genes of different ages (Zhou et al. 2008, Zhang et al. 2010b).

Besides the published reviews on the origins of new genes (Long et al. 2003, Zhou & Wang 2008, Kaessmann 2010), additional summaries are needed for other important and fast-developing topics such as the recent advances in our understanding of the functional significance of new genes, the genome-wide patterns of new gene origination, the discovery of the importance of de novo origination, and functional evolution of new genes. In this review, we attempt to summarize the newest advances in our understanding of the origin of novel genes and their new functions. We conclude with a preview of the future directions of this field.

2. MOLECULAR MECHANISMS FOR ORIGINS OF NEW GENES

2.1. Gene Duplication

Various mechanisms can give rise to a new gene (Long et al. 2003, Zhou & Wang 2008, Kaessmann 2010). Among them, gene duplication is probably the most commonly and extensively studied. A preexisting gene can spawn another copy through small-scale events, such as duplication of the complete/partial gene regions or segmental duplication encompassing several genes. A new gene can also result from an event of much larger scale such as polyploidy or whole-genome duplication (WGD).

WGD is a special case of gene duplication (Semon & Wolfe 2007). Analysis of yeast species' genomes before and after WGD proposed that whether a WGD duplicate is retained as a new

gene is largely determined by its dosage effects rather than by the novel functions it has acquired (Wapinski et al. 2007). For example, genes that encode cellular components or participate in essential growth processes tend to display haploinsufficiency and maintain uniform copy numbers after WGD, while genes that encode peripheral transporters or participate in stress responses tend to show both gene gain and loss through WGD (Wapinski et al. 2007). Recent studies in *Arabidopsis thaliana* and *Saccharomyces cerevisiae* have also revealed some distinctive features of new genes derived from WGDs and single-gene duplications: Whereas single-gene duplication is usually involved in distinctive functional categories (Maere et al. 2005), WGD duplicates tend to share more protein interactions and exhibit less profound phenotypic effects than single-gene duplicates (Hakes et al. 2007).

Smaller-scale gene duplication can be mediated by both DNA and RNA. DNA-based duplication, in a broader sense, occurs with genome segments regardless of whether genes are involved within the segment, whereas an RNA-mediated duplication event (also called retroposition; see below) occurs only in complete or partial gene regions. In detail, DNA-based duplication can emerge through the mediation of repetitive elements [nonallelic homologous recombination (NAHR)] (Roth et al. 1985, Bailey et al. 2003) or as a result of replication errors [nonhomologous end joining (NHEJ)] (Roth & Wilson 1988, Koszul et al. 2004). NHEJ seems to have a broader impact and may occur prior to NAHR (**Figure 1*a***) (Hastings et al. 2009). For example, examination of the breakpoints of 53 telomeric duplications revealed that 92% of them are consistent with NHEJ (Linardopoulou et al. 2005). A recent whole-genome scan of DNA copy-number variation in *Drosophila melanogaster* and *Drosophila simulans* also found no association between the duplication hot spots and the repetitive elements, but it did find an enrichment of duplications in late-replicating regions, suggesting NHEJ rather than NAHR underlies the pattern (Cardoso-Moreira et al. 2011). However, these results do not preclude the importance of NAHR to the production of new duplicates. In a laborious search using fluorescence in situ hybridization for species-specific new genes in eight *Drosophila* species, Yang et al. (2008) identified 17 new genes dispersedly located from their parental genes. Most (82%) of them also have repetitive elements located at one or both of their duplication breakpoints. Further comparative genomics studies (Zhou et al. 2008) confirmed NAHR may play a more important role producing dispersed gene duplicates, whereas NHEJ dominates the generation of tandem duplication in *Drosophila*.

2.2. Retroposition

RNA-mediated duplication or retroposition (for a thorough review, see Kaessmann et al. 2009) is distinct from the above-mentioned molecular processes: The new retrocopy is usually intronless, most likely retaining only part of the parental gene and rarely inheriting the parental gene's promoters (**Figure 1*b***). These defects led to the earlier thought that retrogenes are usually processed pseudogenes (Brosius 1991, Jeffs & Ashburner 1991, Petrov et al. 1996, Mighell et al. 2000). But since the late 1980s (McCarrey & Thomas 1987), abundant retrogenes with intriguing novel functions have been reported. They frequently evolve testis-biased gene expression patterns (Betran & Long 2003, Vinckenbosch et al. 2006, Bai et al. 2007) and may participate in spermatogenesis (Marques et al. 2005), brain function (Viale et al. 2000, Rosso et al. 2008), immune defense (Sayah et al. 2004), and courtship behavior (Dai et al. 2008). These exciting case studies from various species have shifted genome-wide inspections of the evolutionary pattern of retrogenes to a broader scale.

So far, retrogenes have been systematically characterized throughout the genomes of plants (Zhang et al. 2005, Wang et al. 2006), fruit flies (Bai et al. 2007, Zhou et al. 2008), and mammals (human, mouse, opossum) (Emerson et al. 2004, Vinckenbosch et al. 2006, Potrzebowski et al.

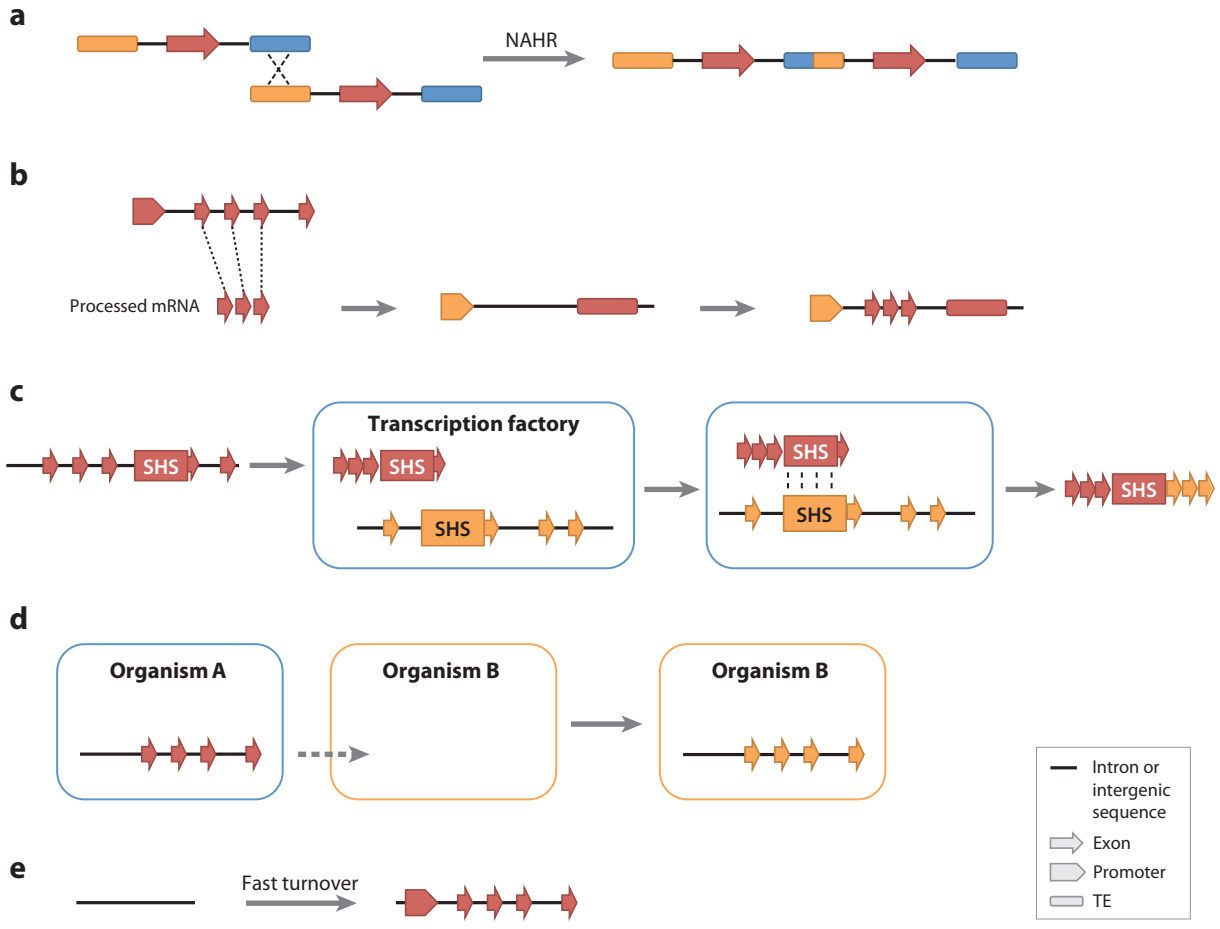

Figure 1

Molecular mechanisms for origins of new genes. (*a*) DNA-level gene duplication produced by nonallelic homologous recombination (NAHR). Orange and blue bars represent nonhomologous transposable elements (TEs). The recombination between them produces a tandem duplication of the focal gene (*red arrows*). (*b*) Retroposition followed by exon shuffling. The reverse-transcription products of the parental genes (*red arrows*) are inserted into a new locus and then recruit nearby sequences as promoters (*orange arrows*). Exon shuffling can happen via the evolution of chimeric gene structure formed by events such as the recruitment of TE sequences into the coding region (*red bar*). (*c*) RNA-level exon shuffling. Two DNA templates share the same transcription factory (TF) for processing mRNAs. Occasionally, the TF containing the processed RNA of one gene (*red arrow*) switches to another gene via the pairing of short homologous sequences (SHSs) without finishing the transcription of the first gene. The following transcription process on another gene creates a chimeric RNA. (*d*) Horizontal gene transfer. Genes of organism A (*red arrows*) may occasionally be transferred to organism B. (*e*) De novo origination. Some noncoding sequences undergo fast turnover and become a functional gene.

2008). General principles regarding the evolution of novel retrogenes, independent of the species being surveyed, have begun to emerge. First, the distinctive testis-biased expression patterns of novel retrogenes do not appear to be restricted to limited cases. For example, a search for species (group)-specific retrogenes using 12 *Drosophila* genomes found that 53% of retrogenes were expressed in the testis and that the retrogenes also tended to be expressed in fewer tissues than were their parental copies (Bai et al. 2007). Similarly, 50 out of 120 identified human retrogenes have evolved testis-related functions (Vinckenbosch et al. 2006). In particular, for both fruit flies

and eutherian mammals, the testis-biased expression pattern for the retrogenes of X chromosome origin is more striking than that of autosomal origin (Betran et al. 2002, Emerson et al. 2004, Potrzebowski et al. 2008). In addition to their regulatory and expression innovations, retrogenes frequently recruit new exons from transposable elements or the sequences of another gene at the inserted site to form a chimeric structure. Investigators estimated that 42% (380 out of 898) of the rice retrogenes (Wang et al. 2006) and 31% (36 out of 117) of the human retrogenes (Vinckenbosch et al. 2006) have formed such a structure, suggesting the innovation of a protein function. These results together indicate that, because retrogenes will insert into a new genomic locus that has a distinctive chromatin status and adjacent sequence feature, the new genomic environment may endow the retrogenes with a new promoter, exon sequences, or chromatin-expression context to evolve novel functions. Interestingly, a recent scan of imprinted genes in the mouse genome has characterized four retrogenes derived from the X chromosome. They are all embedded within another host gene and evolved new methylation patterns during oogenesis (Wood et al. 2007). Although the molecular details of the evolution of imprinting on these retrogenes remain to be disclosed, this case represents a good example of how retrogenes acquire new epigenetic features following their insertion into different genomic loci (Wood et al. 2007).

2.3. Exon Shuffling and *Trans*-Splicing

The idea that exon shuffling would have great impact on protein evolution and new gene formation can be traced back to when the intron was discovered (Gilbert 1978). Intron-mediated recombination was hypothesized to reshuffle exons of different genes and thus create new genes (Patthy 1999). Broadly speaking, the shuffling process could also occur through DNA-level duplication and subsequent recruitment via processes other than intron recombination, such as formation of the chimeric structure after retroposition (**Figure 1*b***). At the RNA level, chimeric RNAs of different gene sources can also form through *trans*-splicing or transcription slippage (**Figure 1*c***) (Li et al. 2009).

The best new gene example of DNA-level exon shuffling may be the case of *jingwei* (Long & Langley 1993), which was the first characterized young gene. This gene was so labeled in reference to a princess in a Chinese myth who is reincarnated into a beautiful bird after death, so as to indicate the rebirth and reformation of the retrogene (Long & Langley 1993). This gene is found only in the *Drosophila yakuba* lineage and is only several million years old. Its emergence started from a retroposition inserted into another duplicated gene *yande*. The retrocopy then recruited three exons of *yande* and formed the new gene *jingwei*. It also evolved novel functions participating in hormone/pheromone metabolism different from its parental gene *Adh* (Zhang et al. 2004). DNA-level exon shuffling can also occur by recruiting intergenic sequences into new exons after gene duplication, such as the case of *Hun* (Arguello et al. 2006), or by recruiting both exon and intron sequences to form novel spliced RNA genes, as in the case of *sphinx* (Wang et al. 2002). The domestication of transposable elements can also result in DNA-level exon shuffling, as in the case of *SETMAR* (Cordaux et al. 2006). A systematic examination of *Drosophila* new genes confirmed a high frequency (30%) of chimeric genes among new genes and showed that various genomic sources can provide materials for exon shuffling (Zhou et al. 2008).

Chimeric genes can also form at the RNA level by *trans*-splicing, which fuses partial pre-RNAs of two distant genes during RNA processing (Gingeras 2009). Early examples of chimeric RNA and their protein products were detected in plants (Chapdelaine & Bonen 1991), insects (Horiuchi et al. 2003, Robertson et al. 2007), and normal (Li et al. 2008) and cancerous (Eychène et al. 2008) human cells, some of which were characterized with differential expression patterns (Li et al. 2008) or conserved across distant species (Gabler et al. 2005). Later genome-wide searches

in model organisms using expressed sequence tags or next-generation sequencing illustrate that *trans*-splicing may be a universal phenomenon increasing transcriptome diversities (Li et al. 2009, McManus et al. 2010, Zhang et al. 2010a), although many of these chimeric RNAs may be derived from artificial chimeric cDNAs (McManus et al. 2010).

In addition, a novel mechanism producing chimeric RNAs starts to emerge (**Figure 1*c***). For example, a recent examination of junction sequences between two source genes within a large pool of chimeric RNAs did not produce canonical GU-AG splicing sites, inconsistent with the splicing model producing the chimeras. Instead, short homologous sequences appear frequently at the junctions and seem to mediate the production of the chimeric RNAs: Mutations in these sequences could abolish the corresponding chimeric RNAs in yeast (Li et al. 2009). Li et al. (2009) thus proposed a transcription slippage to account for the generation of these chimeric RNAs. Nevertheless, the functions of these chimeric RNAs remain largely uncharacterized, and some of the chimeric RNAs have recently been proposed to be experimental artifacts (Houseley & Tollervey 2010). Further studies are needed to clarify this new genetic phenomenon.

2.4. Horizontal Gene Transfer

Horizontal gene transfer (HGT) describes the nonsexual movement of genetic information between different species or between organelles and nuclei (**Figure 1*d***) (Keeling & Palmer 2008). It occurs frequently among prokaryotes (Beiko et al. 2005) and microbial eukaryotes (Gladyshev et al. 2008) and between host and parasitic plants (Yoshida et al. 2010) but rarely in other eukaryotic lineages, probably because the separation of germ and soma sets up a barrier to HGT in most eukaryotes. One of the most striking examples of bacteria-to-animal HGT may be the case of *Wolbachia*, which is an intracellular endosymbiont parasitic to a wide range of arthropods and filarial nematodes (Stouthamer et al. 1999). Its genome sequences have been identified in a wide variety of its host species, and in an extreme case, the entire *Wolbachia* genome (~1.4 Mb) was integrated into the *Drosophila ananassae* nuclear genome with at least 28 transferred genes transcribed (Hotopp et al. 2007). However, no solid case has yet shown the functionality of HGT genes in a multicellular eukaryote. Thus, its contribution to the origin of novel functional genes in multicellular eukaryotes remains to be confirmed.

2.5. De Novo Origination

Initially, the probability that noncoding sequences undergo fast turnover and become a functional new gene (**Figure 1*e***) was thought to be very low (Ohno 1970, Jacob 1977). However, a growing resource of genome sequences has enabled researchers to search for such de novo–originated genes and for their origination rate. The results indicate that de novo origination played an important role during new gene evolution, and case studies of de novo genes further confirmed their functional importance, despite their young ages. Below, we describe and discuss more about this long-underappreciated mechanism of gene origination.

3. GENOME-WIDE PATTERNS OF THE ORIGIN OF NEW GENES

3.1. The Rate of New Gene Origination

Because the majority of new genes are derived from gene duplication or retroposition, our current knowledge about the genome-wide tempo of new gene origination comes mainly from studies of gene duplication rates. As a main cause of spontaneous mutations, the rate of gene duplication

Table 1 Gene duplication rates of different species[a]

Species	Inference from recent duplicates	Inference from mutation accumulation experiments
Saccharomyces cerevisiae	$1 \times 10^{-5} - 8 \times 10^{-3}$	5.9×10^{3}
Caenorhabditis elegans	0.0208	10
Drosophila melanogaster	0.001–0.0023	NA
Homo sapiens	0.00369	NA

[a]Duplication rates are counted as duplicates per gene per million years. The conversion between units of different studies was made assuming 4.8 cell cycles per day for *S. cerevisiae*, and 4 days as the generation time for *C. elegans*. Data were collected or converted from Lipinski et al. (2011), Lynch & Conery (2000), Lynch et al. (2008), and Zhou et al. (2008). Abbreviation: NA, not available.

is of great interest and essential to understanding the impact of duplication on genetic variation. There are two main methods of calculating gene duplication rates: via inference from recent duplicates in sequenced genomes (Lynch & Conery 2000) or directly estimated from mutation-accumulating lines. The current estimates in yeast and worm show the latter will be several orders of magnitude higher than the former (**Table 1**) (Lynch et al. 2008, Lipinski et al. 2011). Such a drastic difference may be derived from differences in the quantification method; more importantly, inferences from recent genome duplicates may suffer from the effects of selection and gene conversion. Nevertheless, these data show the rate of gene duplication in different species is sufficiently large to spawn new genes.

A recent study in the *Drosophila* genome (Zhou et al. 2008) focused on the rate of origination of functional new genes. Because newly nascent genes may still be in the processes of accumulating deleterious/adaptive mutations and often show copy-number polymorphism within the population, this study investigated only functional new genes, which are of greater biological significance and are already fixed within a species. Zhou et al. (2008) studied the new genes that originated in a common ancestor of three *Drosophila* species. Such new genes have recent origins (aged 5–10 million years) and can be found intact in multiple species. The origination rate estimated from these genes is approximately three to five new genes per genome per million years, which is twofold lower than the rate from species-specific new genes.

3.2. A Dynamic View of New Gene Evolution over Time

Comparisons between new genes of different ages uncovered their dynamic patterns over time. For example, characterization of *Drosophila* species-specific new genes showed tandem duplication dominates the production of very nascent copies (~82%), but not of those "older" new genes that are shared by several species (whose proportion drastically decreased to ~34%). Instead, dispersed duplication that separates the new gene copy from its parental gene accounts for 44% of the new genes. These results indicate that individual mechanisms contribute differently to new genes of different ages and that dispersed duplication is more likely to allow for the functional diversification between the new gene and its parental gene (Zhou et al. 2008).

The chromosomal distribution of new genes also changed over time, particularly for male-biased new genes. Studies in mammalian (Zhang et al. 2010c) and *Drosophila* genomes (Zhang et al. 2010b) consistently found an excess of young male-biased protein-coding/microRNA genes that originated on the X chromosome, but this pattern reversed through time, i.e., older male-biased genes that originated earlier are enriched on autosomes and are deficient on the X chromosome. This striking switch is probably a result of several antagonistic factors

simultaneously acting on the X chromosome: The fast-X effect tends to fix an excess of recessive male-beneficial new genes on the X chromosome, whereas meiotic sex chromosome inactivation or sexual antagonism demasculinize the X chromosome and relocate the male-biased genes to autosomes.

4. ORIGINS OF NOVEL NONCODING RNA GENES

Noncoding RNA (ncRNA) genes compose a specific group of genes that function without being translated into proteins; they include the well-documented transfer RNA (tRNA) and ribosomal RNA (rRNA), many kinds of small nonmessenger RNA (snmRNA) genes, as well as the long noncoding RNA (lncRNA) genes (Eddy 2001). They are widespread and play critical roles in regulating various biological processes (Mattick & Makunin 2006, Mercer et al. 2009). The origin of ncRNA had not received much attention until recently.

The snmRNA molecules consist of many different functional categories: small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), small interfering RNA (siRNA), microRNA (miRNA), and piwi-interacting RNA (piRNA) (Eddy 2001). Among these, the repertories of miRNA genes are in constant flux, with high birth and death rates during evolution (Fahlgren et al. 2007, Lu et al. 2008, Nozawa et al. 2012). In plants, the transcription of self-complementary RNA structures by inverted gene duplication is a very common mechanism for the origin of miRNA (Allen et al. 2004, Rajagopalan et al. 2006, Fahlgren et al. 2007, Fahlgren et al. 2010). In contrast, de novo emergence of RNA hairpins seems to be a major route for miRNA genesis in animals (Axtell et al. 2011). A study in *Drosophila* species found that none of the newly identified miRNAs were originated by inverted duplication of preexisting miRNA (Lu et al. 2008). Other mechanisms for miRNA genesis in plants and animals include evolution from transposable elements such as miniature inverted-repeat transposable elements (Smalheiser & Torvik 2005, Piriyapongsa et al. 2007, Piriyapongsa & Jordan 2008) and duplication of preexisting miRNA genes with subsequent divergence (Zhang et al. 2007, 2008; Nozawa et al. 2010, 2012). The evolution dynamics of other snmRNA categories has been much less explored but may involve processes similar to those of miRNA genes. For example, the expansion of piRNA in rodents was reported to occur via repetitive element-mediated duplication (Assis & Kondrashov 2009), a process similar to *Alu*-mediated expansion of miRNA in primates (Zhang et al. 2008). More studies of these snmRNAs will provide further insights into their origination patterns and mechanisms.

For lncRNA genes, our current knowledge of their origination processes is based mostly on case studies. The first characterized young lncRNA gene *sphinx* was identified specifically in *D. melanogaster* (Wang et al. 2002). This gene originated through the insertion of a retroposed copy of the ATP synthase chain F gene with subsequent recruitments of nearby exons and introns to form a chimeric gene structure. The newly formed *sphinx* gene experienced rapid sequence turnover and evolved sex-specific alternative splicing. It does not encode proteins, and, surprisingly, abolishment of splicing sites at its recruited intron in the male-specific transcript leads to male-male courtship behavior. This indicates a novel role for *sphinx* in enhancing the heterosexual courtship of *D. melanogaster* (Dai et al. 2008). A similar mechanism was also found for the well-known lncRNA gene *Xist*, which is the key initiator of X chromosome inactivation in eutherian mammals (Duret et al. 2006). Interestingly, in addition to protein-coding sequences, lncRNA genes can also evolve from previously unexpressed noncoding sequences. An example is the *Poldi* gene in house mice, which arose within the past 2.5–3.5 million years from a stretch of intergenic sequences with preexisting cryptic signals for transcript regulation and processing (Heinen et al. 2009). These cases indicate that lncRNA genes can originate from both protein-coding and "junk" DNA sequences.

The ongoing intensive investigation on lncRNAs may discover more novel lncRNA genes and the genome-wide pattern of their origination.

5. THE IMPORTANT ROLE OF DE NOVO GENE ORIGINATION

Complete de novo origination of a protein-coding gene from a noncoding sequence was thought to be nearly impossible: "Each new gene must have arisen from an already existing gene" (Ohno 1970, p. 72). "The probability that a functional protein would appear de novo by random association of amino acids is practically zero" (Jacob 1977, p. 1164). Nevertheless, the idea of de novo origination is particularly fascinating, as it is intrinsically related to the fundamental question of how the protein repertoire evolved to include such enormous diversity (Bornberg-Bauer et al. 2010). Mechanistically, the de novo route is assumed capable of generating proteins that are very different from those encoded by the existing genome. Therefore, it may have a special role in providing genetic materials for drastic and radical functional renovations. Recent characterizations of de novo genes, including both genome-wide analyses and case studies, have evoked an increasing appreciation of the importance of the de novo mechanism in various organisms (Tautz & Domazet-Loso 2011).

The early clues regarding de novo genes were derived from yeast with the completion of the first eukaryotic genome sequence, of which approximately one-third of the identified genes lacked homologues in other lineages (Dujon 1996). However, it was not until the recent availability of genomes of closely related species that the systematic identification of de novo genes has become a possibility, as the use of recently diverged species allows discrimination between de novo origination and rapid gene evolution as well as retrospection of the noncoding history of de novo genes. The first straightforward search for de novo genes at the genome level was performed by Levine et al. (2006) in *Drosophila*. Using the newly available *Drosophila* genomes, they identified five putative *D. melanogaster* and/or *D. simulans*–specific de novo genes that resemble no/poor BLAST hits to the genomes of *Drosophila yakuba*, *Drosophila erecta*, and *Drosophila ananassae*. All five candidates are expressed predominantly in testes. This finding stimulated the subsequent identification of several male-related de novo genes that emerged in *D. yakuba* or in the ancestor of *D. yakuba* and *D. erecta*, using the data of either accessory gland transcriptome (Begun et al. 2006) or testis-derived expressed sequence tags (Begun et al. 2007). For the genome-wide pattern, it was estimated that 5% of the primate orphan genes (Toll-Riera et al. 2009) and 12% of the *Drosophila* new genes originated from noncoding sequences (Zhou et al. 2008), thus casting doubt on the traditional view that de novo origination is extremely rare and indicating a considerable role for de novo mechanisms in creating novel genes. All the above-mentioned studies have adopted a similar criterion in defining de novo genes, i.e., they should not have recognizable homologues in the genomes of their closely related species. Other studies have employed a different criterion requiring the presence of non-protein-coding homologous sequences in the syntenic regions of the out-group species. This strategy led to the identification of 3 putative protein-coding de novo genes in human and 13 in the malaria parasite *Plasmodium vivax* (Yang & Huang 2011). Using both expressional and proteomic evidence, a recent study discovered 60 human-specific de novo genes (Wu et al. 2011).

Given these results together with other case studies, some common features of de novo genes are emerging. For example, these genes are relatively simple in intron/exon structure and tend to encode short and poorly structured proteins (Begun et al. 2006, Levine et al. 2006, Begun et al. 2007, Knowles & McLysaght 2009). Such features are consistent with speculation that de novo genes are nascent genes that evolved from random noncoding sequences and also raise the possibility of an underestimation of de novo origination owing to the potentially incomplete

annotation of genes encoding short peptides; this is supported by the detection of translation signatures of hundreds of short species (group)-specific open reading frames located in nongenic sequences of the *S. cerevisiae* genome (Carvunis et al. 2012). Nevertheless, there are exceptions: Six of the 13 de novo genes identified in *Plasmodium vivax* possess an intron in the gene region (Yang & Huang 2011), and the yeast de novo gene *MDF1* encodes a three-helix-bearing protein (Li et al. 2010b).

Analysis of the yeast de novo gene *BSC4* by Cai et al. (2008) showed that its origination may be a two-step process. The orthologous, but non-protein-coding, loci of the *BSC4* gene in the out-group species are expressed; thus, an ncRNA state might exist between noncoding DNA and protein-coding genes. Similar processes may also occur within the human-specific de novo gene *FLJ33706*, an idea implied by the low expression of its orthologous locus in *Rhesus macaque* (Li et al. 2010a). Considering the abundance of lncRNAs in mammals and other eukaryotes (Mercer et al. 2009, Ponting et al. 2009), the above observations are particularly illuminative in suggesting the potential of lncRNAs to serve as a rich resource for de novo genes. However, de novo genes may also evolve directly from noncoding DNA, as supported by the example of *MDF1*, whose orthologous loci are not expressed in the out-group species (Li et al. 2010b). In its orthologous sequences, the number of "disabling" nucleotides that can abolish the coding of a fairly long open reading frame seems to decrease gradually toward the branch of *S. cerevisiae MDF1*, raising the possibility that a stepwise process could also exist at the protein-coding level.

As more newly evolved de novo genes are identified across various organisms, few receive comprehensive functional characterization. For the young de novo genes, only the case of *MDF1* (Li et al. 2010b) was narrowed down to the molecular pathway and its functional consequences. *MDF1* is a young de novo gene restricted only to the *S. cerevisiae* lineage. However, it regulates the canonical mating pathway and thereby enables the yeast to make the favorable choice between sexual and asexual reproduction under different nutritional environments (see details below). A series of other recent studies has provided clues regarding the functionality of de novo genes in various organisms. These findings include the discovery of the synthetic lethal gene *BSC4* in yeast, the bacterial infection–associated gene *OsDR1* in rice (Xiao et al. 2009), the spermatogenesis-related gene *hydra* (Chen et al. 2007), three pupae-lethal genes in fruit fly (Chen et al. 2010), and a human-specific gene *FLJ33706* that shows elevated expression in the brain samples of Alzheimer's patients (Li et al. 2010a). Despite the fact that these case studies show the involvements of de novo genes in multiple biological processes, they likely have a much more prominent role in male reproduction, as suggested by its strong testis-biased expression pattern (Levine et al. 2006, Zhou et al. 2008, Wu et al. 2011). Except for testes, de novo genes also show high expression in the human cerebral cortex (Wu et al. 2011), indicating that they may also be functionally important for the evolution of human-unique phenotypic traits such as our dramatically increased cognitive ability.

6. HOW NEW GENES EVOLVE NOVEL FUNCTIONS

Previous attempts to address the functional origin of new genes mainly focused on discussions of the functional fate of new genes, leading to the development of a theoretical framework consisting of various empirically supported models (for reviews, see Conant & Wolfe 2008, Innan & Kondrashov 2010), such as the neofunctionalization (Ohno 1970, Walsh 1995) and subfunctionalization (Force et al. 1999, Lynch & Force 2000) models. However, other fundamental issues remained: What are the functional roles of new genes, and how do they integrate into the preexisting pathways of the host organisms to execute functions? Studies of the functions of young genes are emerging to reveal a critical role of new genes in fundamental

developmental processes, and several recent cases have started to shed light on how new genes may have acquired novel functions through pathway integration and their contributions to pathway evolution.

6.1. Essentiality of Newly Evolved Genes

Given the functional characterizations of newly originated genes, there does not seem to be a necessary association of new genes with species-specific traits. Research now indicates that new genes may play critical roles in essential biological processes that are responsible for reproduction and survival of the organisms. For example, the *Drosophila* new gene *K81* is a paternal-effect gene required for the first round of zygotic division (Loppin et al. 2005), *mojoless* is essential for the survival of male germ line (Kalamegham et al. 2007), and a very young duplicate gene in *D. melanogaster*, *nsr*, is essential for sperm maturation. For the genome-wide-level evidence, Chen et al. (2010) identified 566 young *D. melanogaster* genes that originated within 3–35 million years and systematically tested their phenotypic effects by RNA interference. Surprisingly, they found that 30% of these genes are essential for viability, which is comparable to that estimated for all genes in *D. melanogaster* (~25–35%). The functional essentiality of new genes revealed by these studies has challenged the traditional view that critical biological functions are always encoded by evolutionarily conserved genes. More such studies, especially in other organisms besides fruit fly, are needed to clarify the generality of this observation.

6.2. Pathway Integration of Novel Genes

Genes do not function alone. Upon entering the genome, a new gene has to establish interplay with preexisting gene pathways/networks. Even though a number of new genes have been demonstrated to be functionally important, an in-depth understanding regarding the functional origination of new genes at the pathway level has been lacking. Recently, through rigorous efforts, investigators have made available several cases showing how new genes acquired functions through pathway integration (Matsuno et al. 2009, Li et al. 2010b, Ding et al. 2010, Chen et al. 2012). These studies have also opened up the possibility that new genes are significant drivers of invention, adaptive evolution, and/or turnover of genetic pathways.

Gene duplication can provide enzymes with novel catalytic and recognition properties; this role in the evolution of biochemical pathways has been emphasized since the 1960s (Bryson & Vogel 1965, Jensen 1976). Matsuno et al. (2009) reported the first explicit example of how newly originated genes led to the invention of a novel metabolic pathway. They identified *CYP98A8* and *CYP98A9* as a pair of novel *Brassicaceae* P450 genes that arose via retroposition and duplication, and experienced neofunctionalization as a result of selective and local amino acid replacement. The two new genes then recruited novel catalytic substrates, leading to the synthesis of phenolamide and the formation of a novel phenolic pathway.

Li et al. (2010b) provided a detailed example of how a newly originated de novo gene integrated into the upstream position of the canonical mating pathway and facilitated the organism's adaptation (**Figure 2*a***). The *MDF1* gene evolved from the noncoding antisense sequences of the *ADF1* locus in *S. cerevisiae* and encodes an open reading frame of 152 amino acids. It promotes vegetative growth and decreases mating efficiency in rich mediums. In yeast, the mating pathway is controlled by the *MAT* loci that encode the master regulators of cell types: the *MATa1* locus, present in cells and diploids, and the *MATα1/α2* locus, present in α cells and diploids. The mating behavior is triggered when α-specific genes are turned on by the MATα1 protein and α-specific genes are turned

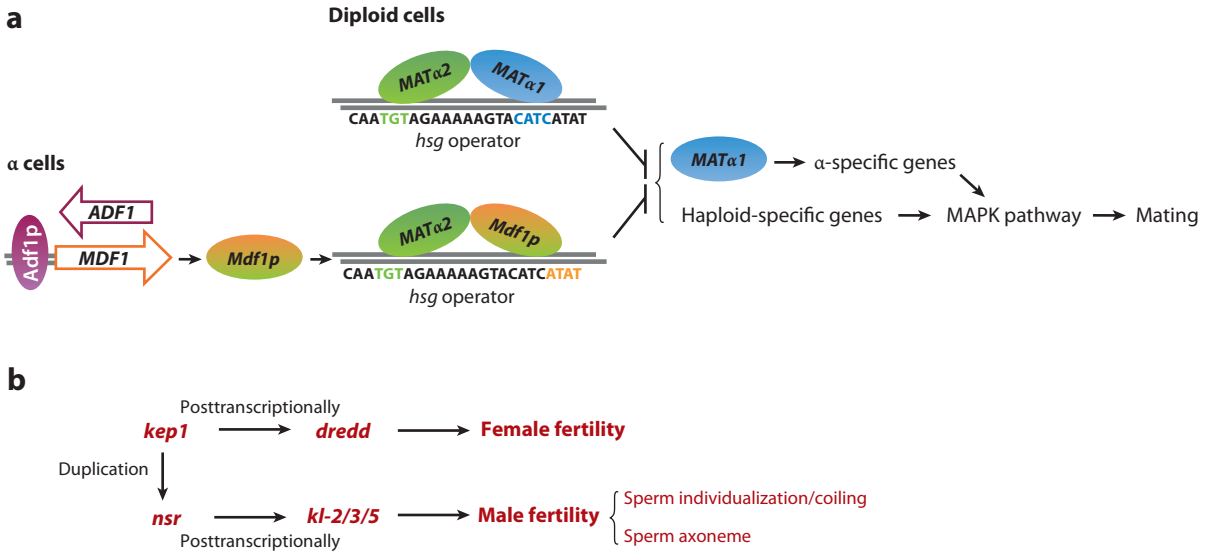

Figure 2

Examples of pathway integration of novel genes. (*a*) The molecular pathway of the de novo new gene *MDF1* in baker's yeast (Li et al. 2010b). In the diploid cell *(upper panel)*, *MATα1* interacts with *MATα2* and then suppresses the downstream α cell-specific and haploid-specific genes. But in the early growing α cell, *Mdf1p* can instead interact with *MATα2* and suppress the similar downstream pathways *(lower panel)*, making the haploid α cell look like a diploid cell and thus avoid mating. (*b*) Pathway integration of the young duplicate gene *nsr* in *Drosophila* (Ding et al. 2010). The parental gene *kep1* takes part in the splicing of *dredd* in females, while its young daughter gene *nsr* perhaps joins in the processing of transcripts of three male-specific genes (*kl-2/3/5*) in testis.

off by the MATα2 protein in α cells. At the early proliferation stage with sufficient nutritional supplement, the MDF1 protein, Mdf1p, binds the MATα2 protein and cooperatively targets the promoters of many haploid-specific genes (*hsg*) and the opener of α-specific genes, *MATα1*. It also downregulates their gene expression levels in α cells. Thus, the MAPK pathway responsible for switching on the physiological changes in preparation for mating is silenced, pushing the yeast toward mitotic cell growth. The mystery of Mdf1p's binding to MATα2 may lie in its structural similarity to MATa1, which binds MATα2 and suppresses mating in diploid cells. However, as a de novo originated protein, Mdf1p is not identical to MATa1: The position of its precise binding sites on the *hsg* operator is four nucleotides away from that of MATa1. Intriguingly, the inhibitory effect of *MDF1* on sexual reproduction is further regulated by its antisense counterpart *ADF1*. The ADF1 protein, Adf1p, can bind to the promoter of *MDF1* and act as a negative regulator. It lets yeast escape from the mitosis cell cycle under unfavorable conditions, such as nutritional limit, to enjoy the benefit of sexual reproduction. Corroborating this, Mdf1p is unable to promote growth in a nonfermentative medium. Therefore, the de novo emergence of the *MDF1* gene has added a new layer of regulatory control to the canonical mating pathway, enabling baker's yeast to adapt well to changing nutritional conditions. The example of *MDF1* shows that newly created genes can regulate a preexisting core pathway in an amazingly delicate manner and has elevated our understanding of the functional contribution of novel genes to organism adaptation.

Two case studies in *Drosophila* show that recently originated genes can rapidly evolve functions and gene networks different from those of their ancestral copy (Ding et al. 2010, Chen et al. 2012). The *D. melanogaster* new gene *nsr* emerged within the past 5.4–12.8 million years through

duplication of the ancestral locus *kep1*. Both *nsr* and *kep1* encode RNA-binding proteins and likely act as splicing factors. The new gene *nsr* is essential for male fertility as it is required for the normal sperm individualization/coiling processes and the structural integrity of the sperm axoneme by regulating the Y chromosome fertility genes *kl-2*, *kl-3*, and *kl-5*, whereas the ancestral gene *kep1* is required for female fertility by regulating the apoptosis molecule *dredd* (**Figure 2*b***) (Ding et al. 2010). *Zeus* is another young male fertility gene, and it originated through retroposition of a highly conserved housing keeping gene, *Caf40*. *Zeus* has a diverged genomic binding profile from *Caf40* and has recruited a new set of downstream genes, which might shape the evolution of gene expression in its germ line (Chen et al. 2012). These cases indicate that the newly derived gene could be quickly incorporated into a pathway that is not analogous to that of its parental copy. It is interesting to note that the sperm individualization process and the structure of the sperm axoneme have been evolutionarily conserved from fly to human. Thus, the case of *nsr* also indicates that the evolution of molecular pathways for essential and conserved biological aspects may be an ongoing process, and new genes may play a crucial role in the adaptive evolution of these molecular pathways. This understanding may help us to explain the "nonorthologous gene displacement" phenomena observed among different organisms (Koonin et al. 1996, Chothia et al. 2003) that a lot of apparently indispensable biological functions are encoded by nonorthologous, and sometimes unrelated, genes. Via this new perspective, the origin of new genes provides the raw material for evolution to play rock and roll.

FUTURE ISSUES

1. The current genome-wide pattern of the origin of new genes was elucidated in *Drosophila* using new genes aged 5–10 million years (Zhou et al. 2008). Does this pattern apply to other species or a different timescale? Beyond the origin of new genes, identifying novel genes in different taxa is also important to understand the evolution of specific traits in a particular taxon. With the rapid advancement of genome sequencing, more reference genomes within phylogenetic contexts are becoming available, and the study of novel genes may be extended to other taxa, such as ants in which the genomes of seven species were recently reported (Gadau et al. 2012). According to the BGI G10K Plant and Animal Reference Genomes Project, many reference genome sequences from a phylogenetic taxon will be deciphered in the near future (**http://ldl.genomics.cn/page/pa-research.jsp**). The Oryza Map Alignment Project will provide a series of reference genomes for rice species with AA genomes (**http://www.omap.org/resources.html**). When these projects are completed, the study of novel genes will proceed with unprecedented breadth and depth.
2. A fast-growing data set is starting to piece together the picture revealing that different genomic components, including noncoding DNA, ncRNA loci/genes, and protein-coding genes, have the potential to "transform" from one to the other. For instance, ncRNA genes can evolve from both noncoding DNA and protein-coding genes. To what extent and scale such events happened at the genome level awaits further elucidation. How our understanding progresses will depend largely on the development and application of advanced proteomics technology and the accumulation of transcriptome and proteomics data regarding closely related species.

3. Though great progress has been made in characterizing the novel functions of new genes, little is known about how they acquired these functions through pathway/network integration. Thus, rigorous investigation could focus on the functions of these new genes at the pathway level, and several recent cases could provide paradigm examples for such explorations. One unanticipated, but increasingly appreciated, finding is that a lot of new genes play key roles in very basic developmental processes, implying a significant contribution of new genes to the evolutionary turnover of fundamental pathways. Further efforts are needed to help researchers explain how closely related species without a certain new gene manage to fulfill the related developmental process and wherein lies the adaptive significance of the pathway turnover driven by the new genes.
4. In recent years, synthetic biology has opened a promising direction in biotechnology that may revolutionize many manufacturing sections with the potential of contributing significantly to human society in an environmentally friendly manner. The study of novel genes and the evolution of their functions can shed insight into the two main aspects of synthetic biology, i.e., constructing artificial biological systems and synthesizing genes or, perhaps, complete genomes (Nielsen & Keasling 2011). Analogous to exon shuffling in nature, artificially shuffling domains of different proteins has successfully rewired pathways, thereby changing the morphology and behavior of cells (Yeh et al. 2007, Peisajovich et al. 2010). After revealing many different kinds of exon/domain-shuffling of novel genes, we may get some ideas of how to design domain recombinants. Alien metabolic pathways have also been introduced into chassis host cells to produce drug precursors (Ro et al. 2006, Westfall et al. 2012). Understanding how novel gene products integrate into or potentially recruit pathways/networks in nature will provide clues valuable for the engineering of manufacture pathways. Nowadays large genomes can be chemically synthesized (Gibson et al. 2010); we believe comprehensive knowledge regarding the origins of novel genes and the evolution of their functions will help researchers design artificial genes, including ones that do not exist in nature, to fulfill human demands.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holding that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Y.D. and Q.Z. contributed equally to this review. All the authors thank Drs. Huifeng Jiang, Guojie Zhang, and Yang Dong for providing useful information. This work was supported by the National Natural Science Foundation of China (30930056, 30623007)and the Chinese 973 program (2007CB815700) to W.W.

Erratum

LITERATURE CITED

Allen E, Xie Z, Gustafson AM, Sung GH, Spatafora JW, Carrington JC. 2004. Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. *Nat. Genet.* 36:1282–90

Arguello JR, Chen Y, Yang S, Wang W, Long M. 2006. Origination of an X-linked testes chimeric gene by illegitimate recombination in *Drosophila*. *PLoS Genet.* 2:e77

Assis R, Kondrashov AS. 2009. Rapid repetitive element-mediated expansion of piRNA clusters in mammalian evolution. *Proc. Natl. Acad. Sci. USA* 106:7079–82
Axtell MJ, Westholm JO, Lai EC. 2011. Vive la différence: biogenesis and evolution of microRNAs in plants and animals. *Genome Biol.* 12:221
Bai Y, Casola C, Feschotte C, Betran E. 2007. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in *Drosophila*. *Genome Biol.* 8:R11
Bailey JA, Liu G, Eichler EE. 2003. An *Alu* transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* 73:823–34
Begun DJ, Lindfors HA, Kern AD, Jones CD. 2007. Evidence for de novo evolution of testis-expressed genes in the *Drosophila yakuba/Drosophila erecta* clade. *Genetics* 176:1131–37
Begun DJ, Lindfors HA, Thompson ME, Holloway AK. 2006. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. *Genetics* 172:1675–81
Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. *Proc. Natl. Acad. Sci. USA* 102:14332–37
Betran E, Long M. 2003. *Dntf-2r*, a young *Drosophila* retroposed gene with specific male expression under positive Darwinian selection. *Genetics* 164:977–88
Betran E, Thornton K, Long M. 2002. Retroposed new genes out of the X in *Drosophila*. *Genome Res.* 12:1854–59
Bornberg-Bauer E, Huylmans AK, Sikosek T. 2010. How do new proteins arise? *Curr. Opin. Struct. Biol.* 20:390–96
Brosius J. 1991. Retroposons—seeds of evolution. *Science* 251:753
Bryson V, Vogel HJ. 1965. Evolving genes and proteins. *Science* 147:68–71
Cai J, Zhao R, Jiang H, Wang W. 2008. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* 179:487–96
Cardoso-Moreira M, Emerson JJ, Clark AG, Long M. 2011. *Drosophila* duplication hotspots are associated with late-replicating regions of the genome. *PLoS Genet.* 7:e1002340
Carvunis A, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, et al. 2012. Proto-genes and de novo gene birth. *Nature* 487:370–74
Chapdelaine Y, Bonen L. 1991. The wheat mitochondrial gene for subunit I of the NADH dehydrogenase complex: a *trans*-splicing model for this gene-in-pieces. *Cell* 65:465–72
Chen S, Ni X, Krinsky BH, Zhang YE, Vibranovski MD, et al. 2012. Reshaping of global gene expression networks and sex-biased gene expression by integration of a young gene. *EMBO J.* 31:2798–809
Chen S, Zhang YE, Long M. 2010. New genes in *Drosophila* quickly become essential. *Science* 330:1682–85
Chen ST, Cheng HC, Barbash DA, Yang HP. 2007. Evolution of hydra, a recently evolved testis-expressed gene with nine alternative first exons in *Drosophila melanogaster*. *PLoS Genet.* 3:e107
Chothia C, Gough J, Vogel C, Teichmann SA. 2003. Evolution of the protein repertoire. *Science* 300:1701–3
Clark AG. 1994. Invasion and maintenance of a gene duplication. *Proc. Natl. Acad. Sci. USA* 91:2950–54
Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nat. Rev. Genet.* 9:938–50
Cordaux R, Udit S, Batzer MA, Feschotte C. 2006. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl. Acad. Sci. USA* 103:8101–6
Dai H, Chen Y, Chen S, Mao Q, Kennedy D, et al. 2008. The evolution of courtship behaviors through the origination of a new gene in *Drosophila*. *Proc. Natl. Acad. Sci. USA* 105:7478–83
Ding Y, Zhao L, Yang S, Jiang Y, Chen Y, et al. 2010. A young *Drosophila* duplicate gene plays essential roles in spermatogenesis by regulating several Y-linked male fertility genes. *PLoS Genet.* 6:e1001255
Dujon B. 1996. The yeast genome project: What did we learn? *Trends Genet.* 12:263–70
Duret L, Chureau C, Samain S, Weissenbach J, Avner P. 2006. The *Xist* RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* 312:1653–55
Eddy SR. 2001. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* 2:919–29
Emerson JJ, Kaessmann H, Betran E, Long M. 2004. Extensive gene traffic on the mammalian X chromosome. *Science* 303:537–40
Eychène A, Rocques N, Pouponnot C. 2008. A new *MAF*ia in cancer. *Nat. Rev. Cancer* 8:683–93

Fahlgren N, Howell MD, Kasschau KD, Chapman EJ, Sullivan CM, et al. 2007. High-throughput sequencing of *Arabidopsis* microRNAs: evidence for frequent birth and death of *MIRNA* genes. *PLoS One* 2:e219

Fahlgren N, Jogdeo S, Kasschau KD, Sullivan CM, Chapman EJ, et al. 2010. MicroRNA gene evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana*. *Plant Cell* 22:1074–89

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–45

Gabler M, Volkmar M, Weinlich S, Herbst A, Dobberthien P, et al. 2005. *Trans*-splicing of the *mod*(*mdg4*) complex locus is conserved between the distantly related species *Drosophila melanogaster* and *D. virilis*. *Genetics* 169:723–36

Gadau J, Helmkampf M, Nygaard S, Roux J, Simola DF, et al. 2012. The genomic impact of 100 million years of social evolution in seven ant species. *Trends Genet.* 28:14–21

Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, et al. 2010. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 329:52–56

Gilbert W. 1978. Why genes in pieces? *Nature* 271:501

Gingeras TR. 2009. Implications of chimaeric non-co-linear transcripts. *Nature* 461:206–11

Gladyshev EA, Meselson M, Arkhipova IR. 2008. Massive horizontal gene transfer in bdelloid rotifers. *Science* 320:1210–13

Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol.* 8:R209

Haldane J. 1935. *The Causes of Evolution*. London: Longmans & Green

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* 10:551–64

Heinen TJ, Staubach F, Haming D, Tautz D. 2009. Emergence of a new gene from an intergenic region. *Curr. Biol.* 19:1527–31

Horiuchi T, Giniger E, Aigaki T. 2003. Alternative *trans*-splicing of constant and variable exons of a *Drosophila* axon guidance gene, *lola*. *Genes Dev.* 17:2496–501

Hotopp JCD, Clark ME, Oliveira DCSG, Foster JM, Fischer P, et al. 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* 317:1753–56

Houseley J, Tollervey D. 2010. Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. *PloS One* 5:e12271

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

Jacob F. 1977. Evolution and tinkering. *Science* 196:1161–66

Jeffs P, Ashburner M. 1991. Processed pseudogenes in *Drosophila*. *Proc. Biol. Sci.* 244:151–59

Jensen RA. 1976. Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* 30:409–25

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20:1313–26

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

Kalamegham R, Sturgill D, Siegfried E, Oliver B. 2007. *Drosophila mojoless*, a retroposed GSK-3, has functionally diverged to acquire an essential role in male fertility. *Mol. Biol. Evol.* 24:732–42

Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.* 9:605–18

Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. *Genome Res.* 19:1752–59

Koonin EV, Mushegian AR, Bork P. 1996. Non-orthologous gene displacement. *Trends Genet.* 12:334–36

Koszul R, Caburet S, Dujon B, Fischer G. 2004. Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J.* 23:234–43

Levine MT, Jones CD, Kern AD, Lindfors HA, 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. USA* 103:9935–39

Li CY, Zhang Y, Wang Z, Cao C, Zhang PW, et al. 2010a. A human-specific de novo protein-coding gene associated with human brain functions. *PLoS Comput. Biol.* 6:e1000734

Li D, Dong Y, Jiang Y, Jiang H, Cai J, Wang W. 2010b. A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell Res.* 20:408–20

Li H, Wang J, Mor G, Sklar J. 2008. A neoplastic gene fusion mimics *trans*-splicing of RNAs in normal human cells. *Science* 321:1357–61
Li X, Zhao L, Jiang H, Wang W. 2009. Short homologous sequences are strongly associated with the generation of chimeric RNAs in eukaryotes. *J. Mol. Evol.* 68:56–65
Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ. 2005. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* 437:94–100
Lipinski KJ, Farslow JC, Fitzpatrick KA, Lynch M, Katju V, Bergthorsson U. 2011. High spontaneous rate of gene duplication in *Caenorhabditis elegans*. *Curr. Biol.* 21:306–10
Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* 4:865–75
Long M, Langley CH. 1993. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. *Science* 260:91–95
Loppin B, Lepetit D, Dorus S, Couble P, Karr TL. 2005. Origin and neofunctionalization of a *Drosophila* paternal effect gene essential for zygote viability. *Curr. Biol.* 15:87–93
Lu J, Shen Y, Wu Q, Kumar S, He B, et al. 2008. The birth and death of microRNA genes in *Drosophila*. *Nat. Genet*. 40:351–55
Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–55
Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154:459–73
Lynch M, Sung W, Morris K, Coffey N, Landry CR, et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad. Sci. USA* 105:9272–77
Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, et al. 2005. Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. USA* 102:5454–59
Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol.* 3:e357
Matsuno M, Compagnon V, Schoch GA, Schmitt M, Debayle D, et al. 2009. Evolution of a novel phenolic pathway for pollen development. *Science* 325:1688–92
Mattick JS, Makunin IV. 2006. Non-coding RNA. *Hum. Mol. Genet.* 15:R17–29
McCarrey JR, Thomas K. 1987. Human testis-specific *PGK* gene lacks introns and possesses characteristics of a processed gene. *Nature* 326:501–5
McManus CJ, Duff MO, Eipper-Mains J, Graveley BR. 2010. Global analysis of *trans*-splicing in *Drosophila*. *Proc. Natl. Acad. Sci. USA* 107:12975–79
Mercer TR, Dinger ME, Mattick JS. 2009. Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* 10:155–59
Mighell AJ, Smith NR, Robinson PA, Markham AF. 2000. Vertebrate pseudogenes. *FEBS Lett.* 468:109–14
Muller HJ. 1935. The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere. *Genetica* 17:237–52
Nielsen J, Keasling JD. 2011. Synergies between synthetic biology and metabolic engineering. *Nat. Biotechnol.* 29:693–95
Nozawa M, Miura S, Nei M. 2010. Origins and evolution of microRNA genes in *Drosophila* species. *Genome Biol. Evol.* 2:180–89
Nozawa M, Miura S, Nei M. 2012. Origins and evolution of microRNA genes in plant species. *Genome Biol. Evol.* 4:230–39
Ohno S. 1970. *Evolution by Gene Duplication*. New York: Springer
Patthy L. 1999. Genome evolution and the evolution of exon-shuffling: a review. *Gene* 238:103–14
Peisajovich SG, Garbarino JE, Wei P, Lim WA. 2010. Rapid diversification of cell signaling phenotypes by modular domain recombination. *Science* 328:368–72
Petrov DA, Lozovskaya ER, Hartl DL. 1996. High intrinsic rate of DNA loss in *Drosophila*. *Nature* 384:346–49
Piriyapongsa J, Jordan IK. 2008. Dual coding of siRNAs and miRNAs by plant transposable elements. *RNA* 14:814–21
Piriyapongsa J, Marino-Ramirez L, Jordan IK. 2007. Origin and evolution of human microRNAs from transposable elements. *Genetics* 176:1323–37

Ponting CP, Oliver PL, Reik W. 2009. Evolution and functions of long noncoding RNAs. *Cell* 136:629–41
Potrzebowski L, Vinckenbosch N, Marques AC, Chalmel F, Jegou B, Kaessmann H. 2008. Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol.* 6:e80
Rajagopalan R, Vaucheret H, Trejo J, Bartel DP. 2006. A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes Dev.* 20:3407–25
Ro DK, Paradise EM, Ouellet M, Fisher KJ, Newman KL, et al. 2006. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* 440:940–43
Robertson HM, Navik JA, Walden KK, Honegger HW. 2007. The bursicon gene in mosquitoes: an unusual example of mRNA *trans*-splicing. *Genetics* 176:1351–53
Rosso L, Marques AC, Weier M, Lambert N, Lambot MA, et al. 2008. Birth and rapid subcellular adaptation of a hominoid-specific CDC14 protein. *PLoS Biol.* 6:e140
Roth DB, Porter TN, Wilson JH. 1985. Mechanisms of nonhomologous recombination in mammalian cells. *Mol. Cell. Biol.* 5:2599–607
Roth DB, Wilson J. 1988. Illegitimate recombination in mammalian cells. In *Genetic Recombination*, ed. R Kucherlapati, GR Simth, pp. 621–53. Washington, DC: Am. Soc. Microbiol.
Sayah DM, Sokolskaja E, Berthoux L, Luban J. 2004. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature* 430:569–73
Semon M, Wolfe KH. 2007. Consequences of genome duplication. *Curr. Opin. Genet. Dev.* 17:505–12
Smalheiser NR, Torvik VI. 2005. Mammalian microRNAs derived from genomic repeats. *Trends Genet.* 21:322–26
Stouthamer R, Breeuwer J, Hurst G. 1999. *Wolbachia pipientis*: microbial manipulator of arthropod reproduction. *Annu. Rev. Microbiol.* 53:71–102
Tautz D, Domazet-Loso T. 2011. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* 12:692–702
Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, et al. 2009. Origin of primate orphan genes: a comparative genomics approach. *Mol. Biol. Evol.* 26:603–12
Viale A, Courseaux A, Presse F, Ortola C, Breton C, et al. 2000. Structure and expression of the variant melanin-concentrating hormone genes: Only PMCHL1 is transcribed in the developing human brain and encodes a putative protein. *Mol. Biol. Evol.* 17:1626–40
Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proc. Natl. Acad. Sci. USA* 103:3220–25
Walsh JB. 1995. How often do duplicated genes evolve new functions? *Genetics* 139:421–28
Wang W, Brunet FG, Nevo E, Long M. 2002. Origin of *sphinx*, a young chimeric RNA gene in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* 99:4448–53
Wang W, Zheng H, Fan C, Li J, Shi J, et al. 2006. High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell* 18:1791–802
Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449:54–61
Westfall PJ, Pitera DJ, Lenihan JR, Eng D, Woolard FX, et al. 2012. Production of amorphadiene in yeast, and its conversion to dihydroartemisinic acid, precursor to the antimalarial agent artemisinin. *Proc. Natl. Acad. Sci. USA* 109:655–56
Wood AJ, Roberts RG, Monk D, Moore GE, Schulz R, Oakey RJ. 2007. A screen for retrotransposed imprinted genes reveals an association between X chromosome homology and maternal germ-line methylation. *PLoS Genet.* 3:e20
Wu DD, Irwin DM, Zhang YP. 2011. De novo origin of human protein-coding genes. *PLoS Genet.* 7:e1002379
Xiao W, Liu H, Li Y, Li X, Xu C, et al. 2009. A rice gene of de novo origin negatively regulates pathogen-induced defense response. *PLoS One* 4:e4603
Yang S, Arguello JR, Li X, Ding Y, Zhou Q, et al. 2008. Repetitive element-mediated recombination as a mechanism for new gene origination in *Drosophila*. *PLoS Genet.* 4:e3
Yang Z, Huang J. 2011. De novo origin of new genes with introns in *Plasmodium vivax*. *FEBS Lett.* 585:641–44
Yeh BJ, Rutigliano RJ, Deb A, Bar-Sagi D, Lim WA. 2007. Rewiring cellular morphology pathways with synthetic guanine nucleotide exchange factors. *Nature* 447:596–600
Yoshida S, Maruyama S, Nozaki H, Shirasu K. 2010. Horizontal gene transfer by the parasitic plant *Striga hermonthica*. *Science* 328:1128

Zhang G, Guo G, Hu X, Zhang Y, Li Q, et al. 2010a. Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res.* 20:646–54

Zhang J, Dean AM, Brunet F, Long M. 2004. Evolving protein functional diversity in new genes of *Drosophila*. *Proc. Natl. Acad. Sci. USA* 101:16246–50

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

Zhang R, Wang YQ, Su B. 2008. Molecular evolution of a primate-specific microRNA family. *Mol. Biol. Evol.* 25:1493–502

Zhang Y, Wu Y, Liu Y, Han B. 2005. Computational identification of 69 retroposons in *Arabidopsis*. *Plant Physiol.* 138:935–48

Zhang YE, Vibranovski MD, Krinsky BH, Long M. 2010b. Age-dependent chromosomal distribution of male-biased genes in *Drosophila*. *Genome Res.* 20:1526–33

Zhang YE, Vibranovski MD, Landback P, Marais GA, Long M. 2010c. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol.* 8:e1000494

Zhou Q, Wang W. 2008. On the origin and evolution of new genes: a genomic and experimental perspective. *J. Genet. Genomics* 35:639–48

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, et al. 2008. On the origin of new genes in *Drosophila*. *Genome Res.* 18:1446–55

Annual Review of Ecology, Evolution, and Systematics

Volume 43, 2012

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Ecology, Evolution, and Systematics* articles may be found at http://ecolsys.annualreviews.org/errata.shtml