Chapter 7

The Origin and Evolution of New Genes

Margarida Cardoso-Moreira and Manyuan Long

Abstract

New genes are a major source of genetic innovation in genomes. However, until recently, understanding how new genes originate and how they evolve was hampered by the lack of appropriate genetic datasets. The advent of the genomic era brought about a revolution in the amount of data available to study new genes. For the first time, decades-old theoretical principles could be tested empirically and novel and unexpected avenues of research opened up. This chapter explores how genomic data can and is being used to study both the origin and evolution of new genes and the surprising discoveries made thus far.

Key words: New genes, Gene duplication, Retrogenes, Gene rearrangements, *De novo* genes, Genetic novelty, Copy number variation

1. Introduction

In the 1940s, geneticists were immersed in a debate over the nature of genetic innovation and organismal complexity (reviewed in ref. 1). The debate centered over determining which class of mutations is responsible for the predominant changes observed between the "primordial" amoeba and men. Are men and amoeba separated only by mutations in preexisting genes or have increases in gene number been a fundamental component of the history of these two lineages? Fifty years onward, we find ourselves in the genomic era, and in possession of the genomes of not only a great number of species, but also of different individuals within the same species. And a comparison of the (several) amoeba and human genomes leaves no doubt as to the origination of new genes being one of the most important sources of evolutionary change.

Most theoretical treatments of the population genetics and molecular evolution of new genes focused on the particular class of gene duplication and preceded the genomic revolution by several decades (e.g., see refs. 2–4). When sequencing technology became

Maria Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods, Volume 2*,
Methods in Molecular Biology, vol. 856, DOI 10.1007/978-1-61779-585-5_7,

readily available in the 1980s, researchers were finally able to empirically study new genes. Initially, only a limited number of new genes were studied in detail, and these were discovered mainly serendipitously (5, 6). In spite of the small sample size, the first examples of new genes began to bring into question long-held views on the mutational processes that generate new genes and on the evolutionary forces that act upon their formation (5, 7). With the onset of the genomic era and the many technologies that it fostered (e.g., in situ hybridization, microarray technology), whole-genome surveys of new genes became feasible. These data allowed researchers to start addressing decades-old questions regarding the early stages of the evolution of new genes. Genome-wide surveys of new genes confirmed several of the previous theoretical predictions and provided a wealth of novel and unexpected observations.

This chapter discusses both the origin and early evolution of new eukaryotic genes, predominantly focusing on the research of the last 10 years that addresses both topics using genome-wide approaches. This chapter is divided into two main sections. The first section explores the different pathways that generate new genes and how the different classes of new genes can be identified from genomic data. The second section focuses on the evolutionary trajectories of new genes. The techniques employed in different studies are described, and the results that are relevant to understanding the evolutionary forces driving the fixation and preservation of new genes in genomes are examined.

2. Origin of New Genes

2.1. Mechanisms of New Gene Origination

New genes are created by a variety of molecular processes, and not all of them are present or are equally active in all genomes. Different molecular pathways generate different classes of new genes, each with distinct molecular signatures that can be recognized from genomic sequence data. Different strategies can be used to date the origin of a new gene, and depending on the class of new gene it might be straightforward or impossible to determine which copy is the original gene (henceforth called parental gene) and which copy is the new gene (henceforth called offspring).

2.1.1. Gene Duplication

Gene duplication is arguably one of the most important sources of evolutionary change and the study of its functional and evolutionary consequences can be traced back to as early as 1911 (1, 8). Duplication events can vary dramatically in size, ranging from a few base pairs to encompassing the complete genome. This review focuses on the smaller class of duplication events, those smaller than a chromosome and larger than a few hundred base pairs, where one or a few new genes are introduced in genomes. Whole-genome duplications

(WGDs) are, however, a very important source of genetic novelties (9), and the readers are encouraged to read Chapter 14, Volume 1 by Kuraku and Meyer (10) of this book, where this phenomenon is discussed. For the purpose of this review, it is important to note that new genes created by small-scale duplications and WGDs differ not only in how they originate, but also in their early evolutionary trajectories. As a consequence, some classes of genes that tend to be fixed after small-scale duplications are not retained in genomes after whole-genome duplication events, and vice versa (9, 11, 12).

As genomes were being sequenced, it became clear that a sizeable portion of all genes (ranging from 17% in some bacteria to 65% in the plant *Arabidopsis*) could be recognized as being duplicates (13). The first whole-genome study of the process of gene duplication was published in 2000 by Lynch and Conery (14) using the then recently fully sequenced fly (*Drosophila melanogaster*), nematode (*Caenorhabditis elegans*), and yeast (*Saccharomyces cerevisiae*) genomes, and the large sequence data already available for the *Arabidopsis* (*A. thaliana*), mouse (*Mus musculus*), and human (*Homo sapiens*) genomes. This was a pioneering study whose methods are still relevant today. Lynch and Conery used gapped BLAST on all translated open reading frames to indentify similar sequence pairs within each genome. They then produced nucleotide sequence alignments for all gene pairs and from them they estimated the fraction of synonymous nucleotide substitutions. Assuming a molecular clock (see Chapter 4, Volume 1 of this book; ref. 112), d_S (i.e., divergence at synonymous sites) can be used as a crude estimate of the age of a duplication. Lynch and Conery calculated the rate of gene duplication using the following data: (1) number of highly similar gene pairs ($d_S < 0.01$, i.e., divergence lower than 1%); (2) estimated number of genes in each of the genomes; and (3) independent estimates of the amount of time needed for two duplicated genes to attain a divergence of 1%. The authors estimated the rate of gene duplication to be between 0.002 (for *Drosophila*) and 0.02 (for the nematode) per gene, per million years. These results were unexpected because they suggested a high rate of gene duplication, on the same order of magnitude as the mutation rate for nucleotide substitutions. With these same data, they also estimated the rate of duplicate gene loss. Lynch and Conery reasoned that if genes are created at a constant rate and if there is no gene loss, when the youngest duplicates ($d_S < 0.25$) are binned into different values of d_S, one should find a similar number of genes in each bin. However, if there is gene loss, one would find instead a decreasing number of genes with increasing d_S. Lynch and Conery found evidence for pervasive gene loss, with more than 90% of gene duplicates disappearing from genomes after only 50 million years, providing an average half-life of 3–7 million years (14).

One limitation of this analysis is its reliance on the molecular clock to estimate the ages of gene duplicates. Although it is reasonable to use the molecular clock to make sequence data comparisons between two species, the model may not hold for duplicate genes as a result of gene conversion (for more details on gene conversion, see Chapter 2, Volume 1 by Budd (15) in this book). If gene conversion is relatively common, older duplicates will falsely appear to be young (reduced d_S), thereby leading to an overestimation of the rate of gene duplication (which is calculated using the number of very young genes ($d_S < 0.001$)). An alternative and more reliable method to the molecular clock is to use a species phylogeny and parsimony to assign gene duplication events to the intervals between the nodes of the phylogenetic tree. Whole-genome sequence data across a species phylogeny became available in 2003 with the published genome sequences of six of *S. cerevisiae* relatives (16, 17). Using these data, Gao and Innan (18) recalculated the age distribution of gene duplicates (originated from WGD and small-scale duplications) using the species tree and arrived at a rate of gene duplication two orders of magnitude lower than the one reported by Lynch and Conery (14). The discrepancy between the two studies suggests that gene conversion plays an important role in the evolution of gene duplicates in yeast genomes, and consequently that the phylogenetic approach is more reliable than relying on the molecular clock (18). In 2007, whole-genome sequence data became available for 12 *Drosophila* species (19), providing a second opportunity to estimate the rate of gene duplication without resorting to the molecular clock. The results of this analysis have, however, been inconclusive. Using the data for all 12 genomes, Hahn and colleagues (20) estimated the rate of gene duplication to be similar to the one calculated by Lynch and Conery (21), thereby suggesting that gene conversion plays a minor role on gene duplicates across the *Drosophila* phylogeny. However, Osaka and Innan (22), using the same data for the *D. melanogaster* subgroup (which corresponds to 4 of the 12 species), arrived at a lower estimate for the rate of gene duplication (but to a lesser degree than the difference found for the yeast genomes), and further found evidence for widespread gene conversion among recent gene duplicates. Despite the disagreement between these two studies on the importance of gene conversion for the evolution of gene duplicates in *Drosophila*, the phylogenetic approach should be robust to the effects of gene conversion and consequently should be favored if the necessary data is available. Another advantage of the phylogenetic approach is that it also avoids the problem of variation in the evolutionary rate at synonymous sites that can also affect the dating of duplicate genes (23).

Duplication events do not have to be restricted to single genes, and quite often encompass multiple genes. As a result, it makes sense to search for the complete stretch of DNA sequence that was duplicated (segmental duplication) instead of only searching for individual gene duplicates. There are two main advantages to this approach: (1) the rate of gene duplication is not overestimated by a single duplication event being counted multiple times and (2) information is gathered on the molecular pathways that generated that mutation. The identification of segmental duplications can also be carried out using the BLAST suite of programs (or similar algorithms). However, instead of using individual gene sequences (amino acid and/or nucleotides), an all-by-all nucleotide genome comparison is required, usually followed by filtering steps aimed at distinguishing duplication events from transposable element sequences, microsatellites, and other repeats (for an example, see ref. 24).

Additional challenges are faced in the detection of duplications that are still polymorphic or that were only recently fixed. These very young duplications have diverged so little between each other that they can be collapsed together when genomes are assembled. As a consequence, the number of very young duplicates may be underestimated from most current genome assemblies. Bailey and colleagues (25) showed that this was an appreciable concern in the human genome by estimating that at least 5% of the human genome is composed of segmental duplications. Bailey and colleagues cleverly reasoned that if they mapped the available whole-genome shotgun reads against the reference genome sequence, the regions that correspond to collapsed segmental duplications should show an increase in read depth resulting from paralogous reads aligning to the same region. Read depth can be calculated using sliding windows along chromosomes, and after segmental duplications are detected their breakpoints can be refined using small-sized windows around the predicted breakpoints (25). This strategy has proven to be relatively successful in identifying segmental duplications in several mammalian genomes (26) and is now routinely used to detect polymorphic duplications using next-generation sequencing data (e.g., see ref. 27). A main caveat of this approach is that the genomic location of the extra copies cannot be retrieved from the analysis.

Determining which of the duplicate copies is the parental gene and which is the offspring can be difficult (Fig. 1). For dispersed duplicates (located distantly from each other), the parent–offspring relationship can be established by combining phylogenetic and syntenic information (Fig. 1). For tandem duplications accompanied by inversions, phylogenetic information combined with gene orientation can also determine the parent–offspring relationship. However, for tandem gene duplications, it may be impossible to distinguish which copy is the parental gene and which copy is the offspring.

There are two main sources of large duplications (and deletions): the imperfect repair of DNA double-strand breaks and DNA replication errors (28). Multiple cellular processes can generate DNA double-strand breaks (e.g., oxidative stress, replication), and since these are highly pathogenic they have to be readily repaired (29). Cells use two main DNA repair pathways to

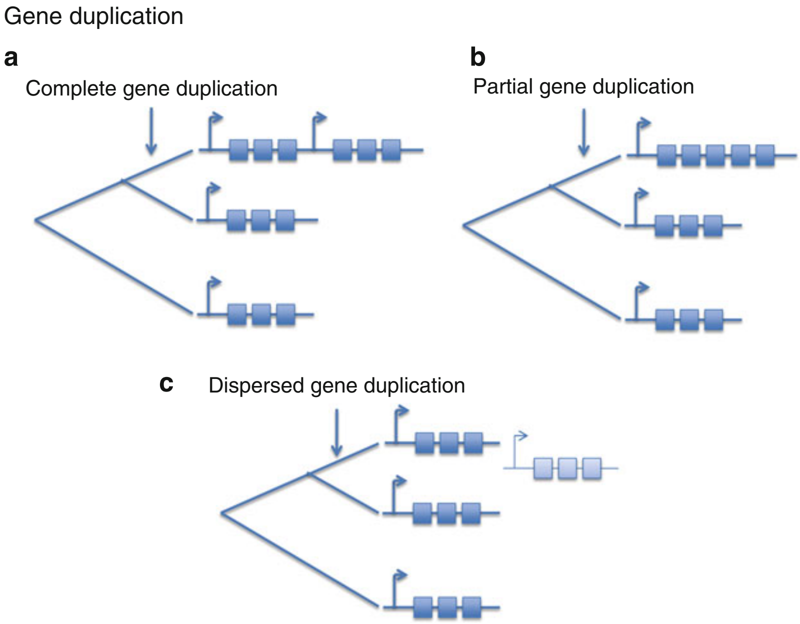

Fig. 1. Schematic depiction of (**a**) complete, (**b**) partial, and (**c**) dispersed gene duplication events as seen in a phylogenetic context. Please note that for complete and partial tandem duplications (**a** and **b**) it may be impossible to distinguish the ancestral from the derived copies. In the case of dispersed duplications (**c**), the parent–offspring relationship can be inferred by combining phylogenetic and syntenic information.

fix these breaks, one that is homology dependent (homologous recombination or HR) and another that is homology independent (nonhomologous end joining or NHEJ) (29, 30). Both HR and NHEJ have been implicated in creating copy number changes (i.e., duplications and deletions). HR can generate duplications (and deletions) when the repair utilizes nonallelic sequences of high sequence identity (instead of the corresponding allele in the sister chromatid or in the homologous chromosome) in a process known as nonallelic homologous recombination (NAHR) (28, 30). Transposable elements, segmental duplications (older duplications already fixed in the species), and other classes of repeats can all mediate NAHR (28, 30). As a result, for young duplications, the role of NAHR can be inferred directly by determining if the duplicated region is flanked by sequences of high sequence identity. In the absence of these sequences, NHEJ or DNA replication errors are assumed to be the underlying mechanism. It has been proposed that DNA replication errors underlie the more complex class of rearrangements (i.e., regions exhibiting multiple structural variants) but it is currently unknown what is its contribution to the formation of simple duplications (and deletions) (28, 30).

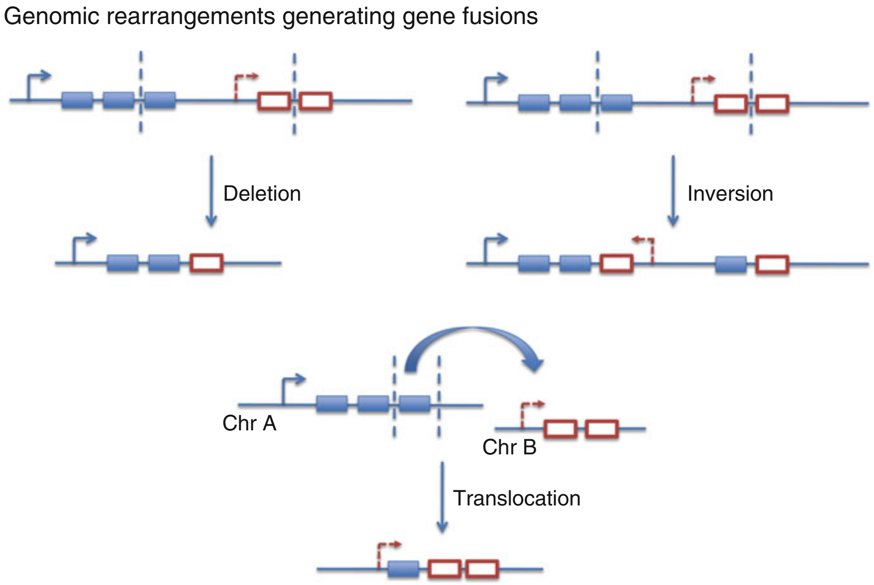

Fig. 2. Schematic depiction of how different classes of genomic rearrangements (deletions, inversions, and translocations) can create fusion genes by juxtaposing sequences from two previously independent genes. All these rearrangements can be preceded by a duplication event, which would allow the creation of a new gene without disrupting the parental genes. The dashed lines represent the area that is mutated (deleted, inversed, or translocated to another genomic location). All examples would create a novel chimeric gene structure.

2.1.2. Genomic Rearrangements

Inversions, translocations, and deletions all have the potential to create new genes by juxtaposing the sequences of two previously independent genes. One example is gene fusion, where two previously distinct genes are fused together in the same transcript creating a novel protein (Fig. 2). Although gene fusions may not be a dominant source of new genes in natural populations (though there are several known examples (31)), they play an important role in many human cancers as gain-of-function mutations (32). Another example of joining distinct genic sequences is exon shuffling, which, as the name suggests, corresponds to recombination-mediated rearrangement of exons between different genes. Exon shuffling is likely to play a major role in the formation of novel protein domains (33, 34). If a duplication precedes the genomic rearrangement, a new gene can be formed while maintaining the parental gene intact. This is expected to increase the probability of the new gene not being deleterious, thus increasing its probability of being fixed.

2.1.3. Retroposition

Retroposition is a class of gene duplication (often called RNA-level duplication or retroduplication) with many distinctive features that distinguish it from the classical model of gene duplication and so merits independent consideration. Retrogenes are created when a

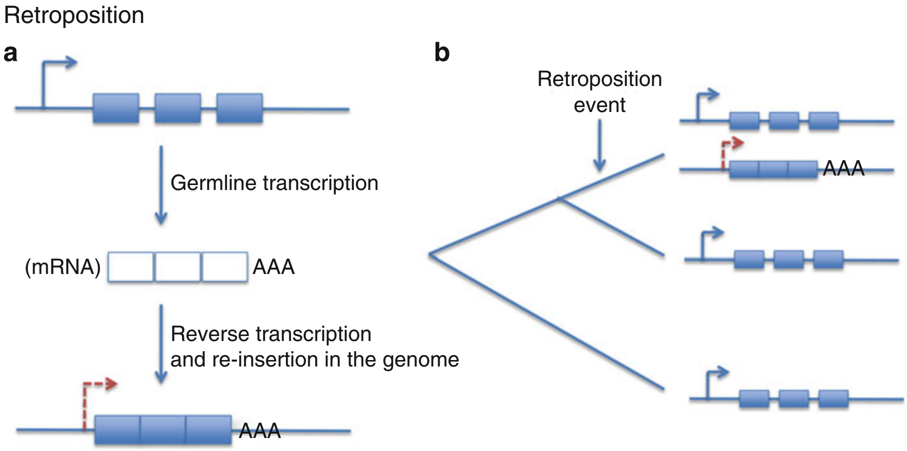

Fig. 3. Schematic representation of how retrogenes are created (**a**) and how they can be identified using a phylogenetic approach (**b**). In (**a**), a retrogene is created after the messenger RNA from the parental gene, intronless and containing a poly-A tail, is reinserted back into the genome. A new regulatory element is then recruited by the new retrogene. A retroposition event can be clearly identified and dated using phylogenetic information (**b**).

messenger RNA is reverse transcribed and inserted back into the genome. Retrogenes are readily identifiable in genome sequences due to several clear hallmarks: (1) absence of introns, (2) presence of a poly-A tail, and (3) flanking short direct repeats. The direct repeats and poly-A tail may not be detectable for older retrogenes, but the presence in a genome of two duplicate genes, one with introns and the other intronless, strongly suggests that the latter was created by retroposition (Fig. 3). The ease with which retrogenes and their parental genes are identified in whole-genome data has made them a model system with which to study new gene formation and evolution (5, 35).

Using the different dating strategies highlighted above, the rates of functional retrogene formation have been estimated for the fly, human, and rice genomes to be of 0.5, 1, and 17 new retrogenes per million years, respectively (36–38). However, retrogenes are not present in all genomes, and if present their abundance can vary greatly between organisms. This is because in order for retroposition to occur two important conditions have to be met: (1) the genome has to possess enzymes capable of reverse transcribing messenger RNAs and integrating the cDNAs back into the genome and (2) those enzymes have to be active in the germ line (in order for retrogenes to be heritable). This may help explain why while the fly and mammalian genomes are very rich in retrogenes, the nonmammalian vertebrate genomes sequenced so far seem to be lacking them (35, 39).

An important feature of retroposition is that it frequently (though not always) generates new genes without regulatory elements. For this reason, retroposition was long believed to be

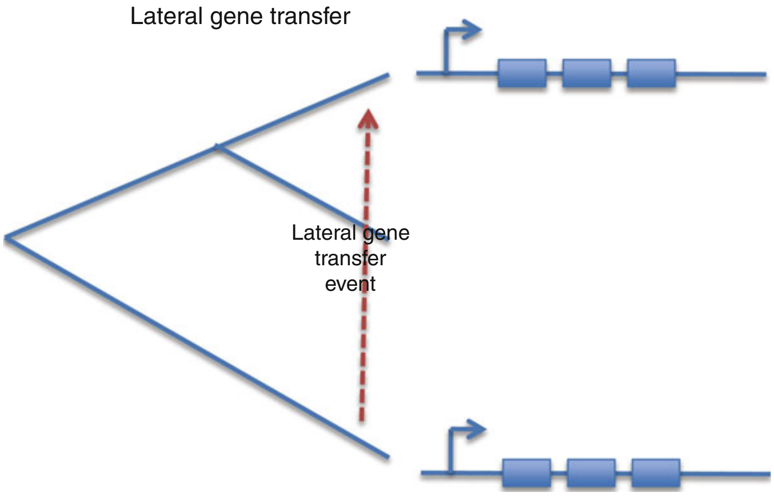

Fig. 4. In a lateral gene transfer event, a gene present in a species is horizontally transferred to another species creating a situation, where the gene tree disagrees with the known species tree.

inconsequential for the origin of new genes. However, a growing number of studies are demonstrating that there are vast numbers of functional retrogenes and that they have been able to recruit regulatory elements through several means (35). For example, retrogenes are often inserted either within or nearby other genes, allowing them to share their regulatory machinery. They can also recruit regulatory elements from nearby retrotransposons, from CpG dinucleotides, as well as evolving *de novo* regulatory elements. Finally, when retrogenes are created from genes with multiple transcript start sites, regulatory elements from the parental gene are also part of the newly formed retrogene (35).

2.1.4. Lateral Gene Transfer

Lateral (or horizontal) gene transfer occurs when a gene is transferred between different organisms (as opposed to being vertically transmitted through the germ line). The laterally transferred gene and its ortholog in the parental lineage are often called xenologs (40). Lateral gene transfer has been shown to be rampant among certain prokaryotic taxa, where it is associated with gains of new genes with many distinct novel functions that contribute dramatically to the evolution of those taxa (41, 42). Lateral gene transfer events can be recognized from genome sequence data in several ways. A lateral gene transfer event generates anomalous or incongruent phylogenetic trees, whereby a given gene may share the highest sequence similarity with a gene in a distantly related species (Fig. 4). Without resorting to phylogenetic trees, genes that have been laterally transferred can be identified in genomes when there are contigs (or sequence reads) that contain sequences readily identified as belonging to different genomes (for example, the presence of both bacterial and eukaryotic gene sequences in the genome of an eukaryote) (43). See Chapter 10, Volume 1 by Lawrence and Azad (44) in this book for more details on how to detect lateral gene transfer events.

Although prokaryote–prokaryote lateral gene transfers are considered to be fairly abundant, prokaryote–eukaryote (and eukaryote–prokaryote) are believed to be much more rare and eukaryote–eukaryote even more. Noteworthy examples of lateral gene transfers between prokaryotes and eukaryotes are the several genes in eukaryotic nuclear genomes that originated from the mitochondrial and plastid genomes (45). Several examples of lateral gene transfers between the bacterial endosymbiont *Wolbachia* and several insect and nematode species have also been documented (prokaryote–eukaryote lateral gene transfer) (46) as have lateral gene transfers from eukaryotes to prokaryotes (47). A remarkable example of lateral gene transfer was found in the pea aphid (*Acyrthosiphon pisum*) genome. When this genome was sequenced in 2010, the authors detected more than ten events of lateral gene transfer from bacteria to this eukaryotic genome (48). However, a limitation to the study's design was that it identified laterally transferred genes from bacterial origin only. Intriguingly, a subsequent study demonstrated that aphids get their orange and red colorations from a set of genes created by duplication events that followed an initial lateral gene transfer from the genome of a fungus (eukaryote–eukaryote lateral gene transfer) (49). The detection of laterally transferred genes should become easier as more sequence data from many different groups of organisms is obtained. These data should also make it possible to quantify the extent of lateral gene transfer between different taxa.

2.1.5. De Novo Gene Origination

De novo genes refer to events, where a coding region originates from a previously noncoding region. *De novo* genes were thought for a very long time to be, at most, rare, even though it was acknowledged that new exons could possibly be added this way (i.e., *de novo* exons*)* (5). However, in 2006, Levine and colleagues (50) reported the existence of five new genes in the *D. melanogaster* genome, all derived from noncoding DNA. This exciting observation was confirmed in subsequent studies on the origin of new genes in *Drosophila* (51, 52) and by discoveries of *de novo* genes in several other genomes (53–56). In order for a new gene to be classified as a *de novo* gene, the orthologous noncoding region in the genome of a close relative should be identified. This is required to show that indeed coding sequence evolved from a previously noncoding sequence (Fig. 5). The presence of a gene in a genome and its absence in the genomes of close relatives does not necessarily imply that that gene evolved *de novo*. For example, that gene could have been lost from all other genomes or it could still be present in those genomes but in regions that are hard to sequence and/or assemble (e.g., heterochromatic regions).

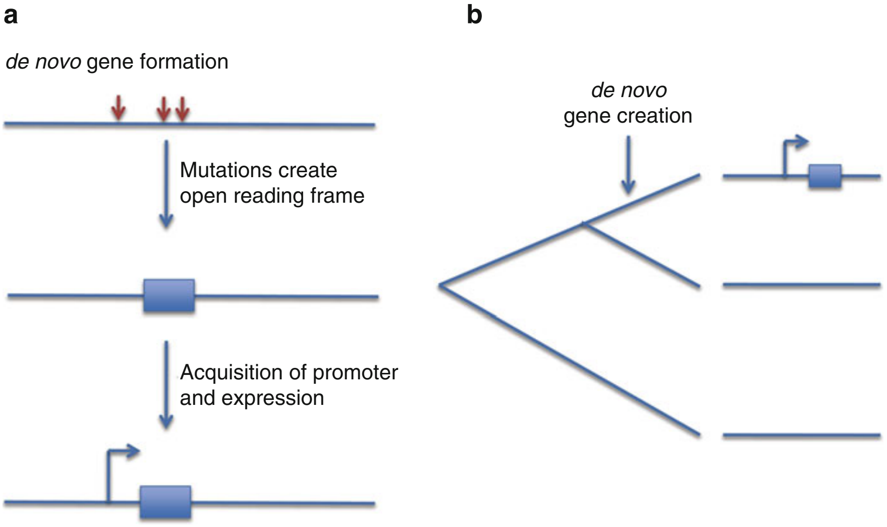
a
de novo gene formation
Mutations create open reading frame
Acquisition of promoter and expression
b
de novo gene creation

Fig. 5. A gene can be created *de novo* when mutations generate a new open reading frame and new regulatory sequences (**a**). Although a *de novo* gene will only be present in the lineage where it was created, orthologous noncoding sequences will be present in closely related taxa (**b**).

2.2. New Noncoding Genes

The repertoire of genes in genomes is not limited to protein-coding genes, but also includes several classes of noncoding RNA genes, such as microRNAs, Piwi-interacting RNAs, and long noncoding RNAs. However, the origin and evolution of noncoding RNA genes are still poorly understood. This reflects the fact that these classes of genes were unknown until recently, but also that they are difficult to detect and present significant challenges for testing functionality.

The first studies aimed at investigating the origin of new noncoding genes focused on gene duplication. These studies revealed an important role for gene duplication in generating microRNAs (57) and Piwi-interacting RNAs (58). However, evidence is still lacking for the role of gene duplication in the formation of long noncoding RNAs (59). Intriguingly, studies of individual long noncoding RNAs, such as the *Xist* in mammals and *spx* in flies, showed that these were created from protein-coding genes, suggesting that this could be a potentially important pathway for the formation of this class of genes (60, 61). Transposable elements are often involved in the formation of new genes either by mediating duplication events (e.g., see ref. 62), by being incorporated into new protein-coding genes as exons, and/or by providing the enzymes needed for retroposition to occur (35). They may play an even more important role in the origination of new noncoding genes as several small RNA genes seem to have emerged from transposable elements (63, 64) as well as some long noncoding RNAs (65). The study of the origin and evolution of novel noncoding genes will likely flourish in the next couple of years, propelled by a better understanding of the molecular biology of these genes.

2.3. Evidence of Functionality in New Genes

The term "new gene" is not indiscriminately applied to any type of novel coding sequence. It is reserved for those gene structures that show evidence of functionality. By definition, a new gene should have an open reading frame, free of any disabling mutations, such as premature stop codons or frameshift mutations. It is important to note, however, that the presence of "disabling" mutations is only suggestive of the absence of functionality. For example, after a gene duplication event, the occurrence of a mutation that shortens the size of the coded protein could potentially generate a new functional protein. More informative is determining if a new gene is evolving under selective constraint (as expected if that gene is functional) or if it is evolving neutrally (as expected from nonfunctional sequences). Information on the selective forces acting on gene pairs can be gathered by determining the rate of synonymous nucleotide substitutions (d_S) and the rate of nonsynonymous (i.e., amino acid replacement) substitutions (d_N) per site. d_N/d_S ratios are commonly calculated between orthologous genes, where a d_N/d_S ratio significantly smaller than 1 suggests that the gene pair is under purifying selection while a d_N/d_S ratio close to 1 suggests that the genes are evolving under no or very little constraint. A third possibility is a d_N/d_S ratio significantly higher than 1, which is suggestive of positive selection (66). *See* Chapter 5 by Kosiol and Anisimova (67) in this volume for details on estimating d_N/d_S. A similar test can be applied to paralogs with a small change. If the parental gene is evolving under functional constraint but the offspring is evolving under no constraint, the d_N/d_S ratio will be significantly smaller than 1 but greater than 0.5 (68). Hence, for new genes, evidence of constraint using a d_N/d_S ratio should conservatively require it to be smaller than 0.5 instead of simply 1 because only the former guarantees that the offspring gene is also under purifying selection. In addition to tests of evolutionary constraint, evidence for transcription and translation of the novel coding sequence provides strong evidence that a putative new gene is functional. However, it is important to note that evidence that a novel coding sequence is expressed is not enough to infer functionality because often bona fide pseudogenes are transcribed (69). Evidence that the new gene is actually translated into a protein constitutes much stronger evidence of functionality (52). Ideally, inferring that a new gene is functional should require several lines of evidence. Moreover, particular classes of new genes may require additional or different lines of evidence to show evidence of functionality, as is the case with *de novo* genes and new noncoding genes.

2.4. Lessons from Genome-Wide Surveys of New Genes

Zhou and colleagues (51) generated the first comprehensive survey of all classes of recently generated new genes for the *D. melanogaster* species subgroup (which comprises four *Drosophila* species). By taking advantage of the 12 *Drosophila* genomes, their well-known phylogeny, and estimated divergence times, they detected all novel genes generated after the split of the *D. melanogaster* species subgroup and dated each event (51). Both sequence similarity and syntenic information were used to infer orthology. Zhou and colleagues found that tandem gene duplications correspond to the vast majority (~80%) of new lineage-specific genes (i.e., genes present in only one species). However, they found a different pattern for older new genes (those shared by multiple species and more likely to be functional): 44% are dispersed gene duplicates (i.e., located distantly from each other) while only 34% occur as tandem duplications. Ten percent of the remaining new genes were created by retroposition and a surprisingly twelve percent were created *de novo*. No lateral gene transfers were detected. Using this subset of older new genes, Zhou and colleagues estimated the rate of new gene origination to range between 0.0004 and 0.0009 per gene per million of years, which translates into 5–11 new genes added to the *Drosophila* genome every million years.

One of the most surprising results coming from surveys of new genes in different genomes is the large amount of chimeric gene structures found. A new gene is considered chimeric if it recruits novel sequence from nearby regions. For example, retrogenes are expected to recruit novel regulatory sequences as the transposition event often leads to the loss of all regulatory sequences from the parental gene. Similarly, gene fusions and exon shuffling generate chimeric gene structures (70). However, gene duplication, which is the mechanism responsible for the creation of most new genes, was thought for a long time to generate two fully redundant copies of a gene (4). As discussed in the next section, population genetic models of the evolution of gene duplicates usually assume this to be the case.

The highest rate of new chimeric gene formation was observed in grass genomes (37), where 7 chimeric genes are fixed every million years, a rate 50 times higher than the one found for humans (36). In the survey of new genes in *Drosophila* mentioned above, Zhou and colleagues (51) found that only 41% of new genes specific to *D. melanogaster* have their coding sequence completely duplicated and that this percentage is even lower for older new genes (16%). They also found that ~30% of all new genes recruit additional flanking sequence. Previous studies on new genes created by gene duplication in the nematode *C. elegan*s also suggested that as much as 50% of all new genes have recruited novel sequences and that most gene duplication events did not encompass the complete gene structure but are instead partial gene duplications (71, 72). Better insight into the mutational processes generating new genes can be gained by looking at the youngest class of all new genes that are still segregating as polymorphisms. Surveys of polymorphic duplications and deletions in both flies and humans (collectively called copy number variants, or CNVs) found that most duplications are indeed partial, with only a minority encompassing complete genes (73, 74).

By comparing new genes of different ages, insight can be gained into the characteristics that increase their probability of being preserved in genomes. Both the distance between the two copies of a gene and the recruitment of other genomic sequences (i.e., creation of chimeric gene structures) seem to increase the probability of a new gene being preserved in a genome for a longer period of time (75).

When knowledge of new genes was limited to individual case studies, two patterns began to emerge. The first was that many new genes were found on the X chromosome. The second was that most new genes were proposed to have male-biased functions, with evidence coming from both expression and functional data (e.g., see refs. 76, 77). Genome-wide surveys of new genes emphatically confirmed both patterns (51). They further showed that this pattern was true for even less conventional classes of new genes, such as *de novo* genes (50, 78). Recent studies have shown that both the distribution of new genes among chromosomes and their expression patterns are dynamic processes. In both fly and mammalian genomes, the youngest class of new genes is enriched on the X chromosome and exhibits male-biased expression (52, 79). However, for older classes of new genes, both patterns change: these genes are less likely to reside on the X chromosome and to have male-biased functions (52, 79). One explanation is that new genes with male-biased expression move progressively through time out of the X chromosome and into the autosomes, leading to an overall paucity of male-biased genes on the X chromosome (52, 79–81). The movement of new genes out of the X chromosome and into the autosomes was first described in *Drosophila* for retrogenes (77), later confirmed in the mouse and human genomes (68) and further shown to also be true for genes created by gene duplication (82). More work is required to determine what is the actual proportion of retrogenes (and new genes in general) that are formed on the X chromosome and then are translocated to the autosomes (50). Global analysis of gene expression of both parental and offspring genes in flies and mammals suggests that meiotic X chromosome inactivation is one of the driving forces behind the movement of new male-biased genes away from the X chromosome (83, 84).

3. The Evolutionary Trajectories of New Genes

Just like any other mutation, new genes can be neutral, deleterious, or advantageous. Except in populations with an extremely small population size, if a new gene is deleterious it will be kept at low frequency in the population, never reaching fixation (i.e., never becoming present in all individuals of the species). Examples of deleterious new genes are duplications of dosage-sensitive genes, where the new copy of the gene leads to a deleterious change of gene expression (85). If a new gene is neutral or advantageous, then it has a chance of becoming fixed. The probability of fixation and the time to fixation depend on the strength of selection. The higher the selective advantage, the likelier it is for the new gene to be fixed and the shorter the time to fixation. It is important to note that the most likely fate for neutral—and even advantageous—new genes is removal from the populations (86). Once a new gene is fixed, its subsequent evolution dictates its probability of being retained in the genome for long periods of time (86, 87). Three main evolutionary fates have been suggested for new gene duplicates and these can be extended to other classes of new genes. They are discussed in detail below.

3.1. Possible Evolutionary Fates for New Genes

3.1.1. Pseudogenization (Nonfunctionalization)

The most likely outcome for a new gene is to become a pseudogene due to the accumulation of inactivating mutations. It has been estimated that there is one pseudogene for every eight functional genes in the *C. elegans* genome (88) and as much as one pseudogene for every two functional genes in the human genome (89). It is important to emphasize that not all pseudogenes are derived from new genes. Many genes that were functional for long periods of time become pseudogenes because of changes in the evolutionary pressures acting on them. For example, it is thought that the reduced use of olfaction in hominoids contributed to the large percentage of pseudogenes in the family of human olfactory receptors (90). The reason that pseudogenization is the most likely outcome for a new gene is that the vast majority of mutations that can occur in a new gene (or in any other genomic sequence) are either neutral or deleterious. Hence, if a new gene is not evolving under constraint, it will sooner or later accumulate enough mutations that render it nonfunctional.

3.1.2. Neofunctionalization

New genes will be preserved in genomes for long periods of time if they confer a novel (advantageous) function. The classical neofunctionalization model advocated by Ohno proposed that after a gene duplication event there would be two redundant copies of the same gene, which would relax selective constraints in one of the copies allowing it to accumulate mutations (4). Although advantageous mutations are rare, if one occurred in one of the copies of the gene it could provide it with a novel function, thereby preserving the new duplicate in the genome. A now classical example of neofunctionalization is the duplication of a pancreatic ribonuclease gene in leaf-eating monkeys. After the duplication, one of the copies evolved rapidly under positive selection for a more efficient digestive function in a new microenvironment (91). Remarkably, this same gene was suggested to have been duplicated independently in Asian and African leaf-eating monkeys and in both monkeys one of the copies evolved under positive selection for more digestive functions (92).

The very large number of duplicates preserved in genomes suggested to some that neofunctionalization could not be responsible for the preservation of all or even of most of them (93, 94). This is because the balance between the number of deleterious and advantageous mutations tilts strongly toward the former. This led different authors to propose alternative models, namely, the different subfunctionalization models described below. However, recent genomic data suggests that novel functions may be more common than previously thought and that they can often be created at the time the new gene is formed. With the exception of complete gene duplications, all other processes that create new genes do not generate two fully redundant copies of the same gene. Partial gene duplications, gene fusions, exon shuffling, retrogenes, and *de novo* genes all create novel gene structures that often recruit nearby genomic sequences. Even if a novel gene structure is not created, the presence of the new gene in a different chromatin environment from its parental gene could potentially already endow it with a new function (e.g., by being able to be expressed under different conditions). Of course, only a small fraction of these novel gene structures are likely to provide a novel function and are thus likely to be fixed and preserved by positive selection (51). Surveys of new genes support the idea that novel gene structures and/or different genomic locations contribute disproportionately to the fraction of new genes that end up being preserved in genomes (51, 62).

3.1.3. Subfunctionalization

The concept that a pair of duplicate genes can share the same function of the ancestral gene is old (1). More recently, this concept has been formalized into distinct models. One of them is called the duplication, degeneration, complementation model (DCC) (93). It posits that after a gene duplication event that generates two fully redundant copies selection is relaxed for both copies and mutations are allowed to accumulate. A mutation that would be deleterious when there was only one copy of the gene is now rendered neutral due to the presence of the other copy. This allows both copies to accumulate degenerative and complementary mutations, which result in the two genes being necessary to fulfill the functions of the original gene. Importantly, this model of subfunctionalization requires only neutral substitutions (as opposed to beneficial mutations) and applies to the partitioning of functions coded both in protein and regulatory sequences. An alternative subfunctionalization model is called the escape from adaptive conflict (EAC) (9, 94, 95). This model assumes that the original gene is capable of two or more distinct functions that cannot be simultaneously optimized by selection due to pleiotropic effects. Gene duplication would allow each of the copies to perform one of the functions that could now be optimized by positive selection. The DCC and EAC models differ in that in the DCC the mutations that cause the subfunctionalization are explicitly neutral and in the EAC they are adaptive.

Neofunctionalization and subfunctionalization are not mutually exclusive. After a subfunctionalization event that preserves the two duplicates in the genome, an advantageous mutation can still occur and create a novel function in one of the duplicates. Subfunctionalization could greatly increase the probability of neofunctionalization by extending the period of time available for an advantageous mutation to occur (96).

3.2. Methods to Detect the Evolutionary Forces Acting on New Genes

3.2.1. Determining the Selective Forces Responsible for the Fixation of New Genes

Understanding the fixation process of new genes requires either the study of recently fixed new genes or the study of new genes that are still polymorphic in the population. When a new gene is fixed, either by neutral genetic drift or by positive selection, it exhibits reduced levels of polymorphism because all individuals in the population share the same recently originated new gene. However, the degree of reduction of polymorphism in the new gene (and also in the parental gene if they are linked) depends on the strength of selection. The stronger the selection, the lower the levels of polymorphism. Positive selection also leads to reduced levels of polymorphism in the sequences surrounding the new gene, a phenomenon referred to as selective sweep (97). The stronger the selection, the more reduced the levels of polymorphism will be and the larger the area surrounding the new gene that exhibits low levels of polymorphism. After the fixation, patterns of polymorphism in both the new gene and the surrounding sequences return to the levels observed before the mutation event, thereby erasing the signature of the selective force responsible for this process (97). Very few studies to date have addressed the fixation process of new genes (a remarkable exception being 98). This is likely to change in the next few years with the proliferation of population genomic data for different species (e.g., 1,000 genomes project, various *Drosophila* population genomics project, *Arabidopsis* population genomics project).

Polymorphic new genes can also provide important information about the process of fixation of new genes. Surveys of CNVs in different species have already identified several candidates to be under positive selection (73, 74). Evidence comes from analyzing patterns of polymorphism surrounding the CNVs as described above and by looking at population differentiation (99, 100). Most CNV studies so far identify polymorphic duplications but often cannot determine the exact number of new copies, their location, or their actual sequence. As next-generation sequencing methods are more widely applied to detect CNVs, these limitations should disappear and detailed sequence analysis of both the polymorphic duplications and their flanking sequences will be available. CNVs can also help elucidate how often new genes are fixed by positive selection due to changes in gene dosage. The combination of expression data and sequence polymorphism can address this question directly.

3.2.2. Identifying the Evolutionary Fates Responsible for the Retention of New Genes

The different models proposed for the fates of new genes make different predictions regarding the early stages of the evolution of new genes. The neofunctionalization model proposed by Ohno predicts that in a duplicate gene pair one member experiences a period of relaxed constraint, followed by a period of positive selection (after the occurrence of the mutation that confers a new function), while the other member continuously experiences purifying selection (4). According to this model, there should be an asymmetric rate of evolution between the two duplicates. This same asymmetry should also be detected for those new genes whose origination immediately confers a new advantageous function. In this case, there should not be any period of relaxed constraint. Instead, the new genes are expected to be driven to fixation by positive selection, which is expected to continue to act for some period of time. Meanwhile, the parental gene is expected to evolve under purifying selection. New genes that are identical to its parental genes could be immediately favored by positive selection due to changes in gene dosage, as numerous examples have demonstrated (e.g., see refs. 99, 101). When this occurs, the new gene is fixed by positive selection, but in this case both parental and offspring genes are expected to be under purifying selection and exhibit a symmetrical rate of evolution.

The subfunctionalization models do not make clear predictions regarding whether gene duplicates are expected to diverge symmetrically or asymmetrically because the functions of the ancestral gene could potentially be divided equally or unequally between the two duplicates. However, at least in its earlier stages, the DCC model would predict both genes to experience relaxed constraint and during this stage their evolution should be symmetrical. The DCC and EAC models can be distinguished from each other because the latter predicts both parental and offspring genes to experience a period of positive selection.

As mentioned above, subfunctionalization and neofunctionalization are not mutually exclusive. New genes may experience an initial stage of subfunctionalization (DCC model) followed by a period of neofunctionalization. This would be translated into an initial period of evolution under relaxed constraints for both genes followed by a symmetrical or asymmetrical period of evolution under positive selection depending on whether the latter acts on one or both duplicates. Another alternative scenario is the fixation of a duplicate by positive selection for dosage alteration that then subsequently evolves a novel function. This scenario would create an initial period of positive selection driving the duplication to fixation, followed by a period of symmetrical evolution, where both members are under purifying selection, and finally another period of positive selection created by the mutation that confers the novel function. The fact that different scenarios can be hypothesized and that the different models do not make explicit enough assumptions to allow for their clear distinction has hampered our capability of determining what are the dominant modes of evolution for new genes (97, 102).

3.2.3. Detecting the Modes of Selection Acting on Parent–Offspring Gene Pairs

Advantageous mutations capable of conferring a new gene with a new function can occur in both coding and noncoding (regulatory) regions. The different methods available to detect positive selection acting on both types of sequence are reviewed in detail in Chapters 5–6 of this volume (67, 103), and can be readily applied to new genes. One such method is using the d_N/d_S ratio to infer if a gene is evolving under purifying selection, neutrality, or positive selection. As discussed below, this method has been applied extensively to the study of new genes and so it is important to note two of its limitations. First, positive selection is of an episodic nature and is followed by a period of purifying selection that can erase the sequence patterns suggestive of positive selection. Therefore, tests based on the d_N/d_S ratio have more power when applied to young genes. Although several techniques have been proposed to detect signs of positive selection in older parent–offspring pairs (reviewed in ref. 104), it is very hard to distinguish among the different evolutionary scenarios for old genes. Second, positive selection may only act on a small subset of the gene with the remaining sequence evolving under purifying selection. In this case, the d_N/d_S ratio also fails to detect positive selection (104). As described in Chapter 5 of this volume by Kosiol and Anisimova (67), there are different techniques that can be used to detect positive selection acting on a subset of the protein sequence.

Distinguishing between the different models proposed for the early evolution of new genes requires determining if the parental gene and its offspring are evolving symmetrically or asymmetrically. Relative rate tests use an outgroup sequence (i.e., an ortholog in a close species of the parental gene) to determine if one of the genes is evolving at a faster rate (104). A faster rate of evolution in one of the genes is compatible with two scenarios: (1) one of the genes is evolving under relaxed constraints while the other is under purifying selection or (2) one of the genes is evolving under positive selection while the other is under purifying selection. Additional data has to be collected to distinguish between these two scenarios. For older new genes that have already had time to accumulate several additional mutations, polymorphism and divergence data can be combined to show that if that gene was evolving neutrally then inactivating mutations would already have had time to accumulate. In this case, the presence of extensive amino acid changes without disruption of the protein-coding sequence is only compatible with positive selection (and not with relaxed selection). For younger genes, the number of nucleotide substitutions is usually not enough to distinguish between the two scenarios.

Evidence for asymmetrical evolution can also be gathered from expression data. A novel function or a partition of functions among duplicates can be detected at the expression level by comparing the patterns of expression of the parental gene, the offspring, and the ortholog of the parental gene in a closely related species (e.g., see ref. 105). Studying the patterns of evolution of pairs of parent–offspring genes of different ages could provide a dynamic picture of the early stages of the evolution of new genes. However, caution has to be taken when doing this type of comparisons. Certain trends that can emerge from this type of analyses may be due to the differential features of preserved vs. nonpreserved gene pairs instead of reflecting the changes through time experienced by preserved gene pairs (96).

3.2.4. Insights from Genome-Wide Surveys of the Early Evolution of New Genes

The first large-scale surveys on the forces acting on duplicated genes found little evidence for positive selection (14, 106). Lynch and Conery (14) calculated d_N/d_S ratios for pairs of gene duplicates in six eukaryotic genomes and found that the vast majority was under purifying selection. The youngest class of gene duplicates showed signs of being under purifying selection even though they were more likely to tolerate amino acid changes than older genes (which could be a sign of relaxed constraints or positive selection) (14). Kondrashov and colleagues (106) applied the same d_N/d_S approach to gene duplicates in 26 bacterial, 6 archaeal, and 7 eukaryotic genomes and also found purifying selection to be the dominant force. They further used an outgroup sequence to compare the rate of evolution between the two duplicates and found that paralogs typically evolve symmetrically (106). Conant and Wagner (107) used a codon-based model that distinguishes between silent substitutions and amino acid replacements when testing for potential asymmetries in protein sequence divergence. This time, evidence was found supporting asymmetrical evolution for 20–30% of duplicate gene pairs in four different eukaryotic genomes. They also found evidence for relaxed selective constraints in those genes evolving asymmetrically with a minority exhibiting signs of being under positive selection (107). As discussed above, in older duplicates, the earlier signs of a period of asymmetrical evolution may have been obliterated by the subsequent period of purifying selection. Hence, it is noteworthy that when Zhang and colleagues focused on young duplicates in the human genome they found that ~60% were evolving asymmetrically (108).

Since some new genes are identical to their parental genes (i.e., complete tandem gene duplications) while others are not (i.e., retrogenes, dispersed duplicates), it merits asking if the percentage of genes evolving at asymmetrical rates is the same for the two classes of genes. Cusack and Wolfe (109) found that the degree of asymmetry in the rate of evolution is greater for gene pairs where parent and offspring genes differ from each other than for those gene pairs where parent and offspring genes are identical. Han and colleagues (110) found a similar result when studying lineage-specific duplicates in the human, macaque, mouse, and rat genomes. By focusing on very young duplicates, they also aimed at detecting signs of positive selection before it was masked by the purifying selection that follows. Approximately 10% of all lineage-specific genes showed signs of positive selection acting in their protein sequences. Furthermore, they showed that for gene duplicates, where parental and offspring genes are located in different genomic locations, 80% of the time that there was evidence for positive selection it came from the offspring copy. This was true when the offspring was a retrogene or was created by the classical model of gene duplication (110).

When divergence data is combined with polymorphism data, further insight can be gained into the evolutionary forces acting on new genes. More precisely, combining both types of data allows distinguishing between the two scenarios that can cause accelerated rates of protein evolution: relaxation of selective constraints and positive selection. Cai and Petrov (111) combined human polymorphism data with human–chimp divergence data and found strong evidence that the elevated rates of protein evolution found for younger genes are mostly due to relaxed selective constraints and found weaker evidence that younger genes experience adaptive evolution more frequently than older genes.

4. Future Perspectives

It is unquestionable that the wealth of genomic data collected in the past 10 years dramatically changed our understanding of how new genes are created. But more than answering long-standing questions, the genomics revolution brought about a brand new set of questions. Only recently have we learned that new genes could be created *de novo* (50–56) and we are still lacking the proper tools to study how selection acts in this group of genes. Also, now that we know that an important component of genomes are nonprotein-coding genes, we have to devise more sensitive detection techniques in order to detect them and study their evolution. And perhaps the greatest challenge of all, we have to go beyond simply describing the sequence and evolution of new genes and determine the novel functions these genes are coding. Although genomic data helps us determining if a gene is functional or not, determining its actual function requires a multidisciplinary effort that combines genomics and proteomics with a multitude of functional assays.

As more genomes are sequenced, phylogenies will become more and more complete and our capability of detecting new genes, dating them, and understanding how they are formed will increase. As we move from sequencing genomes of different species to sequencing many genomes from the same species, we will be able to combine divergence and polymorphism data on a genome-wide scale and finally be able to better describe the evolutionary forces acting on new genes. We will also move from detecting polymorphic new genes using microarray technology to using next-generation sequencing, and with it we will obtain the detailed sequence information on the new genes, their location, and breakpoint information that we are currently lacking. As genomic data continues to accumulate, so will our understanding of how new genes are formed, how they are fixed in populations, and why they are preserved in genomes.

5. Questions

1. Count the number of genes in the human and chimpanzee genomes. Does the difference suggest the gain or the loss of some genes in one lineage? How can you distinguish between the two possibilities?
2. Imagine the genome sequences of 12 bee species (the phylogeny is known) have just been released. The 12 genomes have been annotated using both experimental and computational approaches. What would be the steps needed to find all lineage-specific genes, i.e., genes present in only one of the species? What genomic hallmarks would you use to distinguish the different classes of new genes?

Acknowledgments

We thank J. Roman Arguello, Maria Vibranovski, three anonymous reviewers, and our editor, Maria Anisimova for comments and critical reading of the manuscript.

References

1. Taylor JS, Raes J (2004) Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet 38:615–643
2. Haldane JBS (1932) The causes of evolution. Princeton Science Library
3. Bridges CB (1936) The Bar 'gene' a duplication. Science 83:210–211
4. Ohno S (1970) Evolution by gene duplication. Springer-Verlag
5. Long M, Betrán E, Thornton K et al (2003) The origin of new genes: glimpses from the young and old. Nat Rev Genet 4:865–875
6. Presgraves DC (2005) Evolutionary genomics: new genes for new jobs. Curr Biol 15:R52–53
7. Long M, Langley CH (1993) Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science 260:91–95

8. Kuwada Y (1911) Meiosis in the pollen mother cells of Zea Mays L. Bot Mag 25:1633
9. Conant GC, Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet 9:938–950
10. Kuraku S, Meyer A (2012) Detection and phylogenetic assessment of conserved synteny derived from whole genome duplications. In: Anisimova M (ed) Evolutionary genomics: statistical and computational methods (volume 1). Methods in Molecular Biology, Springer Science+Business Media New York
11. Wapinski I, Pfeffer A, Friedman N et al (2007) Natural history and evolutionary principles of gene duplication in fungi. Nature 449:54–61
12. Maere S, De Bodt S, Raes J (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A 102:5454–5459
13. Zhang J (2003) Evolution by gene duplication: an update. Trends Eco Evo 18: 292–298
14. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155
15. Budd A (2012) Diversity of genome organization. In: Anisimova M (ed) Evolutionary genomics: statistical and computational methods (volume 1). Methods in Molecular Biology, Springer Science+Business Media New York
16. Cliften P, Sudarsanam P, Desikan A et al (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301:71–76
17. Kellis M, Patterson N, Endrizzi M et al (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241–254.
18. Gao LZ, Innan H (2004) Very low gene duplication rate in the yeast genome. Science 306:1367–1370.
19. Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the Drosophila phylogeny. Nature 450:203–218
20. Hahn MW, Han MV, Han SG (2007) Gene family evolution across 12 Drosophila genomes. PLoS Genet 3:e197
21. Lynch M, Conery JS (2003) The evolutionary demography of duplicate genes.J Struct Funct Genomics 3:35–44
22. Osada N, Innan H (2008) Duplication and gene conversion in the Drosophila melanogaster genome. PLoS Genet 4:e1000305
23. Long M, Thornton K (2001) Gene duplication and evolution. Science 293:1551
24. Fiston-Lavier AS, Anxolabehere D, Quesneville H (2007) A model of segmental duplication formation in Drosophila melanogaster. Genome Res 17:1458–1470
25. Bailey JA, Gu Z, Clark RA et al (2002) Recent segmental duplications in the human genome. Science 297:1003–1007
26. Marques-Bonet T, Girirajan S, Eichler EE (2009) The origins and impact of primate segmental duplications. Trends Genet 25:443–454
27. Medvedev P, Stanciu M, Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. Nat Methods 6:S13-20
28. Gu W, Zhang F, Lupski JR (2008) Mechanisms for human genomic rearrangements. Pathogenetics 1:4
29. Aguilera A, Gómez-González B (2008) Genome instability: a mechanistic view of its causes and consequences. Nat Rev Genet 9:204–217
30. Hastings PJ, Lupski JR, Rosenberg SM et al (2009) Mechanisms of change in gene copy number. Nat Rev Genet 10:551–564
31. Rogers RL, Bedford T, Hartl DL (2009) Formation and longevity of chimeric and duplicate genes in Drosophila melanogaster. Genetics 181:313–322
32. Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. Nature 458:719–724
33. Long M, Rosenberg C, Gilbert W (1995) Intron phase correlations and the evolution of the intron/exon structure of genes. Proc Natl Acad Sci U S A 92:12495–12499
34. Patthy L (1999) Genome evolution and the evolution of exon-shuffling–a review. Gene 238:103–114
35. Kaessmann H, Vinckenbosch N, Long M (2009) RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet 10:19–31
36. Marques AC, Dupanloup I, Vinckenbosch N et al (2005) Emergence of young human genes after a burst of retroposition in primates. PLoS Biol 3:e357
37. Wang W, Zheng H, Fan C et al (2006) High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell 18:1791–1802
38. Bai Y, Casola C, Feschotte C et al (2007) Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila. Genome Biol 8:R11
39. Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. Genome Res 20:1313–1326

40. Patterson C (1988) Homology in classical and molecular biology. Mol Biol Evol 5: 603–625

41. Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304

42. Gogarten JP, Townsend JP (2005) Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol 3:679–687

43. Zhaxybayeva O (2009) Detection and quantitative assessment of horizontal gene transfer. Methods Mol Biol 532:195–213

44. Lawrence J, Azad R (2012) Detecting lateral gene transfer. In: Anisimova M (ed) Evolutionary genomics: statistical and computational methods (volume 1). Methods in Molecular Biology, Springer Science+Business Media New York

45. Martin W, Herrmann RG (1998) Gene transfer from organelles to the nucleus: how much, what happens, and Why? Plant Physiol 118:9–17

46. Dunning Hotopp JC, Clark ME, Oliveira DC et al (2007) Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science 317:1753–1756

47. Doolittle RF, Feng DF, Anderson KL et al (1990) A naturally occurring horizontal gene transfer from a eukaryote to a prokaryote. J Mol Evol 31:383–388

48. The International Aphid Genomics Consortium (2010) Genome Sequence of the Pea Aphid Acyrthosiphon pisum. PLoS Biol 8: e1000313

49. Moran NA, Jarvik T (2010) Lateral transfer of genes from fungi underlies carotenoid production in aphids. Science 328:624–627.

50. Levine MT, Jones CD, Kern AD et al (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 103:9935–9939

51. Zhou Q, Zhang G, Zhang Y et al (2008) On the origin of new genes in Drosophila. Genome Res 18:1446–1455

52. Zhang YE, Vibranovski MD, Krinsky BH et al (2010) Age-dependent chromosomal distribution of male-biased genes in Drosophila. Genome Res 20:1526–1533

53. Cai J, Zhao R, Jiang H et al (2008) *De novo* origination of a new protein-coding gene in Saccharomyces cerevisiae. Genetics 179:487–496

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

55. Toll-Riera M, Bosch N, Bellora N et al (2009) Origin of primate orphan genes: a comparative genomics approach. Mol Biol Evol 26:603–612

56. Xiao W, Liu H, Li Y et al (2009) A rice gene of *de novo* origin negatively regulates pathogen-induced defense response. PLoS One 4:e4603

57. Hertel J, Lindemeyer M, Missal K et al (2006) The expansion of the metazoan microRNA repertoire. BMC Genomics 7:25

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

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

60. Duret L, Chureau C, Samain S et al (2006) The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science 312:1653–1655

61. Wang W, Brunet FG, Nevo E et al (2002) Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. Proc Natl Acad Sci U S A 99:4448–4453

62. Yang S, Arguello JR, Li X et al (2008) Repetitive element-mediated recombination as a mechanism for new gene origination in Drosophila. PLoS Genet 4:e3

63. Smalheiser NR, Torvik VI (2005) Mammalian microRNAs derived from genomic repeats. Trends Genet 21:322–326

64. Piriyapongsa J, Mariño-Ramírez L, Jordan IK (2007) Origin and evolution of human microRNAs from transposable elements. Genetics 176:1323–1337

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

66. Wagner A (2002) Selection and gene duplication: a view from the genome. Genome Biol 3: reviews1012

67. Kosiol C, Anisimova M (2012) Selection in protein coding regions. In: Anisimova M (ed) Evolutionary genomics: statistical and computational methods (volume 1). Methods in Molecular Biology, Springer Science+Business Media New York

68. Emerson JJ, Kaessmann H, Betrán E et al (2004) Extensive gene traffic on the mammalian X chromosome. Science 303:537–540

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

70. Arguello JR, Fan C, Wang W et al (2007) Origination of chimeric genes through DNA-level recombination. Genome Dyn 3:131–146

71. Katju V, Lynch M (2003) The structure and early evolution of recently arisen gene duplicates in the Caenorhabditis elegans genome. Genetics 165:1793–1803
72. Katju V, Lynch M (2006) On the formation of novel genes by duplication in the Caenorhabditis elegans genome. Mol Biol Evol 23:1056–1067
73. Emerson JJ, Cardoso-Moreira M, Borevitz JO et al (2008) Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. Science 320:1629–1631
74. Conrad DF, Pinto D, Redon R et al (2010) Origins and functional impact of copy number variation in the human genome. Nature 464:704–712
75. Zhou Q, Wang W (2008) On the origin and evolution of new genes–a genomic and experimental perspective. J Genet Genomics 35:639–648
76. Arguello JR, Chen Y, Yang S et al (2006) Origination of an X-linked testes chimeric gene by illegitimate recombination in Drosophila. PLoS Genet 2:e77
77. Betrán E, Thornton K, Long M (2002) Retroposed new genes out of the X in Drosophila. Genome Res 1854–1859
78. Begun DJ, Lindfors HA, Kern AD et al (2007) Evidence for *de novo* evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. Genetics 176:1131–1137
79. Zhang Y, Vibranovski DV, Landback P et al (2010) Chromosomal Redistribution of Male-Biased Genes in Mammalian Evolution with Two Bursts of Gene Gain on the X Chromosome. PLoS Bio 8:e1000494
80. Ranz JM, Castillo-Davis CI, Meiklejohn CD et al (2003) Sex-dependent gene expression and evolution of the Drosophila transcriptome. Science 300:1742–1745
81. Parisi M, Nuttall R, Naiman D et al (2003) Paucity of genes on the Drosophila X chromosome showing male-biased expression. Science 299:697–700
82. Vibranovski MD, Zhang Y, Long M (2009) General gene movement off the X chromosome in the Drosophila genus. Genome Res 19:897–903
83. Vibranovski MD, Lopes HF, Karr TL et al (2009) Stage-specific expression profiling of Drosophila spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic relocation of testis-expressed genes. PLoS Genet 5:e1000731
84. Potrzebowski L, Vinckenbosch N, Marques AC et al (2008) Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS Biol 6:e80
85. Conrad B, Antonarakis SE (2007) Gene duplication: a drive for phenotypic diversity and cause of human disease. Annu Rev Genomics Hum Genet 8:17–35
86. Otto SP, Yong P (2002) The evolution of gene duplicates. Adv Genet 46:451–483
87. Kondrashov FA, Kondrashov AS (2005) Role of selection in fixation of gene duplications. J Theor Biol 239:141–151
88. Harrison PM, Echols N, Gerstein MB (2001) Digging for dead genes: an analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome. Nucleic Acids Res 29:818–830
89. Harrison PM, Hegyi H, Balasubramanian S et al (2002) Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. Genome Res 12:272–280
90. Rouquier S, Blancher A, Giorgi D (2000) The olfactory receptor gene repertoire in primates and mouse: evidence for reduction of the functional fraction in primates. Proc Natl Acad Sci U S A 97:2870–2874
91. Zhang J, Zhang YP, Rosenberg HF (2002) Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat Genet 30:411–415
92. Zhang J (2006) Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. Nat Genet 38:819–823
93. Force A, Lynch M, Pickett FB et al (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545
94. Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proc Biol Sci 256:119–1124
95. Piatigorsky J, Wistow G (1991) The recruitment of crystallins: new functions precede gene duplication. Science 252:1078–1079
96. Lynch M, Katju V (2004) The altered evolutionary trajectories of gene duplicates. Trends Genet 20:544–549
97. Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11:97–108
98. Moore RC, Purugganan MD (2003) The early stages of duplicate gene evolution. Proc Natl Acad Sci U S A 100:15682–15687

99. Perry GH, Dominy NJ, Claw KG et al (2007) Diet and the evolution of human amylase gene copy number variation. Nat Genet 39:1256–1260
100. Schrider DR, Hahn MW (2010) Gene copy-number polymorphism in nature. Proc Biol Sci 277:3213–3221
101. Schmidt JM, Good RT, Appleton B et al (2010) Copy number variation and transposable elements feature in recent, ongoing adaptation at the Cyp6g1 locus. PLoS Genet 6:e1000998
102. Hahn MW (2010) Distinguishing among evolutionary models for the maintenance of gene duplicates. J Hered 100:605–617
103. Zhen Y, Anfolfatto P (2012) Detecting selection on non-coding genomics regions. In: Anisimova M (ed) Evolutionary genomics: statistical and computational methods (volume 1). Methods in Molecular Biology, Springer Science+Business Media New York
104. Raes J, Van de Peer Y (2003) Gene duplication, the evolution of novel gene functions, and detecting functional divergence of duplicates in silico. Appl Bioinformatics 2:91–101
105. Huminiecki L, Wolfe KH (2004) Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res 14:1870–1879
106. Kondrashov FA, Rogozin IB, Wolf YI et al (2002) Selection in the evolution of gene duplications. Genome Biol 3: RESEARCH0008
107. Conant GC, Wagner A (2003) Asymmetric sequence divergence of duplicate genes. Genome Res 13:2052–2058
108. Zhang P, Gu Z, Li WH (2003) Different evolutionary patterns between young duplicate genes in the human genome. Genome Biol 4:R56
109. Cusack BP, Wolfe KH (2007) Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. Mol Biol Evol 24:679–686
110. Han MV, Demuth JP, McGrath CL et al (2009) Adaptive evolution of young gene duplicates in mammals. Genome Res 19:859–867
111. Cai JJ, Petrov DA (2010) Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. Genome Biol Evol 2:393–409
112. Aris-Brosou S, Rodrigue N (2012) The essentials of computational molecular evolution. In: Anisimova M (ed) Evolutionary genomics: statistical and computational methods (volume 1). Methods in Molecular Biology, Springer Science+Business Media New York