Newly evolved genes: Moving from comparative genomics to functional studies in model systems

How important is genetic novelty for species adaptation and diversification?

José M. Ranz[1)*] *and John Parsch*[2)*]

Genes are gained and lost over the course of evolution. A recent study found that over 1,800 new genes have appeared during primate evolution and that an unexpectedly high proportion of these genes are expressed in the human brain. But what are the molecular functions of newly evolved genes and what is their impact on an organism's fitness? The acquisition of new genes may provide a rich source of genetic diversity that fuels evolutionary innovation. Although gene manipulation experiments are not feasible in humans, studies in model organisms, such as *Drosophila melanogaster*, have shown that new genes can quickly become integrated into genetic networks and become essential for survival or fertility. Future studies of new genes, especially chimeric genes, and their functions will help determine the role of genetic novelty in the adaptation and diversification of species.

Keywords:
■ adaptation; gene function; genetic networks

DOI 10.1002/bies.201100177

[1)] Department of Ecology and Evolutionary Biology, University of California Irvine, CA, USA
[2)] Department of Biology II, University of Munich (LMU), Munich, Germany

***Corresponding authors:**
José M. Ranz
E-mail: jranz@uci.edu
John Parsch
E-mail: parsch@bio.lmu.de

Introduction

The number of protein-encoding genes contained within a genome can vary greatly among species, ranging from <200 in the extreme endosymbiont *Carsonella ruddii* [1] to over 40,000 in cultivated rice [2, 3]. Even closely related species can differ in gene number, which implies that a species' gene content may change over evolutionary time. For example, it has been estimated that humans and chimpanzees, which diverged from a common ancestor about six million years ago [4], do not share 6.5% of their 22,000 genes [5]. Some genes are present only in humans, while others are present only in chimpanzees. This difference in gene content is much greater than the average 1.5% divergence in nucleotide sequence between genes shared by the two species. These observations raise some obvious questions, including: How do differences in gene content arise? Where do new genes come from? Do differences in gene content contribute to phenotypic differences between species? A recent paper by Zhang et al. [6] presents evidence that new genes have played an important role in the evolution of higher cognitive functions in humans. In the following, we review the methodology and results of this paper and highlight several other recent studies aimed at uncovering the functional significance and impact on fitness of newly evolved genes in model organisms.

The availability of complete genome sequences from many diverse species has made it possible to identify specific changes in gene content and determine when they occurred [6–8]. A schematic of the approach followed by Zhang et al. [6] is depicted in Fig. 1. Briefly, the linear order of genes within conserved chromosomal segments is compared among a set of species with a known evolutionary relationship. When a difference in gene content is observed, it represents either a gain or loss of a gene at that location. The presence (or absence) of the gene in other species, in conjunction with their evolutionary relationship, can be used to determine whether the gene was gained or lost. If it was gained, it is considered to be a novel

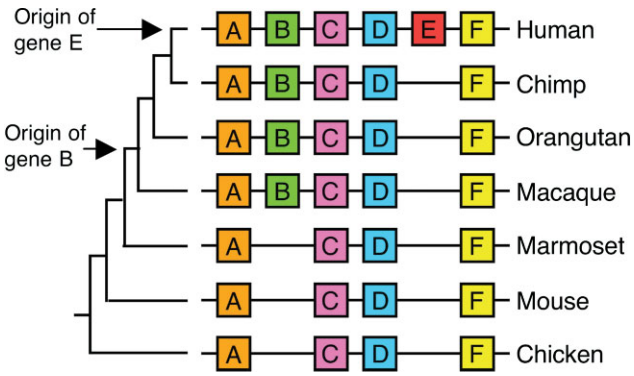


Figure 1. Method to identify new genes. An orthologous chromosomal region is compared among species of known evolutionary relationship and the age of genes (here shown as boxes labeled A–F) is inferred from the lineage of their first appearance. In this example, gene B is a young gene specific to primates, while gene E is newly evolved in the human lineage. Genes A, C, D, and F are old genes that are common to all investigated genomes. This approach has been followed by Zhang et al. [6] and others [5, 7].

gene that is unique to a particular species or lineage and its age can be estimated from the branch on the phylogenetic tree where it first appears. Genes that appeared relatively recently and are shared only by a small number of species are termed ''new'' or ''young'', while those common to many diverse taxa are considered to be ''old'' or ''ancient''. The study of Zhang et al. [6] focused on a set of 1,828 new genes unique to humans and other primates. Using data from expressed sequence tags and microarrays, the authors detected a significant overabundance of these new genes in the brain transcriptome, but not in the transcriptomes of other tissues. Further analyses revealed that the expression of the new genes was particularly enriched in the fetal brain and in the neocortex, which is the brain region thought to be responsible for many human-specific cognitive abilities [9].

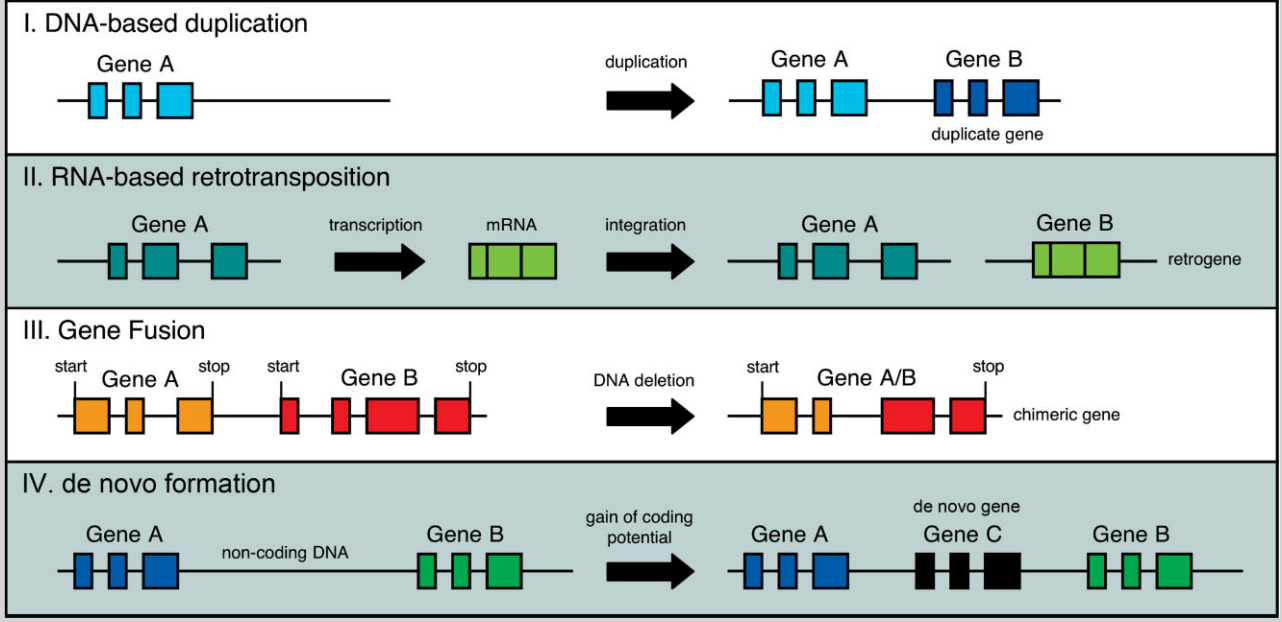
Mechanisms of gene creation

Features of the DNA sequence of a new gene, together with the sequence of the corresponding region in species lacking the gene, can be used to infer the mechanism that led to its creation (Box 1). The most common process that gives rise to new genes is duplication, which can occur either by DNA- or RNA-based mechanisms [10]. The former typically involves unequal crossing-over, which can lead to the expansion in copy number of genes in pre-existing multi-gene families [5, 11]. Unequal crossing over can also occur between interspersed repetitive elements, such as transposable elements, leading to the duplication of single-copy genes and the insertion of duplicate genes on non-homologous chromosomes [12]. RNA-based duplication, or retroduplication, occurs when a transcribed mRNA of an existing gene is reverse transcribed

Box 1

Mechanisms of new gene creation

The following schematic diagrams illustrate the major mechanisms of new gene creation. Exons are depicted as solid boxes, while introns and intergenic regions are represented by lines.



and inserted into a new location in the genome [13, 14]. A new gene can also be created by the fusion of two previously distinct genes to form a chimeric gene, or by the addition or rearrangement of protein-encoding exons within a pre-existing gene [15–22]. It is also possible for new protein-encoding genes to arise de novo from non-coding DNA [23–32]. The creation of new genes by this mechanism is thought to be rare, although recent work suggests that at least 60 de novo genes are unique to the human genome [33].

Evolution after creation: How do new genes evolve?

The selective forces acting on a new gene soon after its formation are likely to vary depending on the mechanism that created the gene. Chimeric and de novo genes represent genetic novelty and can immediately perform new functions that may be subject to natural selection. Gene duplications, on the other hand, lead to genetic redundancy. Initially, a new duplicate gene is expected to be non-essential and evolve under little or no selective constraint. This allows mutations that alter the encoded protein to accumulate. If there is no constraint on the gene and a mutation introduces a stop codon or a frameshift into the coding sequence, it will become a pseudogene and gradually disappear from the genome by recurrent deletion [34]. However, it may be that the relaxed constraint allows the gene to explore the local protein space and the occurrence of chance mutations leads to a protein with a new, advantageous function. If so, the gene will be favored by natural selection and might, over time, become essential for the organism. In both of the above scenarios, the ratio of nonsynonymous to synonymous substitutions (*Ka/Ks*) that occur in the new gene is expected to increase relative to its parental (functional) gene (Box 2). However, only in the latter case will this ratio decrease after a new function has been optimized. If so, the ratio of nonsynonymous to synonymous polymorphism in the gene in the present day population will be less than the ratio of nonsynonymous to synonymous substitutions that occurred between species [35]. The new, primate-specific genes identified by Zhang et al. [6] as being expressed in fetal brain had *Ka*/*Ks* over twice that of old, fetal brain expressed genes or the genome average, which indicates that there has been rapid evolution of new genes expressed during brain development. Nearly one third of these genes (5 out of 16 tested) showed a significant reduction in nonsynonymous polymorphism in the current human population relative to nonsynonymous substitutions that occurred between human and chimpanzee, indicating that the level of selective constraint on these genes has increased since the time of speciation. Zhang et al. [6] interpret the above molecular patterns as evidence for positive selection driving the evolution of new, brain-expressed genes. However, with duplicate genes, similar patterns can be produced by changes in the intensity of negative selection [36].

In addition to changes in protein sequence, new genes can gain functions by acquiring new patterns of expression, which allow them to be active in different tissues or at different stages of development than their parental genes. The acquisition of new expression patterns is thought to be especially important for genes originating through RNA-based duplication, because they insert into regions of the genome far away from their original regulatory sequences [37]. However, novel expression patterns can be gained by genes originated by all of the mechanisms described above. Indeed, Zhang et al. [6] found that all types of newly evolved primate genes showed an enrichment of fetal brain expression, which suggests that gain of expression in the developing brain was a key attribute favoring their retention in the genome.

Determining the functional relevance of newly evolved genes: Model systems offer a way forward

When direct genetic manipulation is not possible, the inference of a new gene's function usually involves the characterization of its expression and the categorization of the type of protein it encodes. Typically, researchers scrutinize databases to gain insights into different components of the expression profile of the gene, namely: abundance, spatial distribution, timing of expression, spliced forms, and the dynamics of expression in the presence of external cues [38–42]. Likewise, the protein's function can be inferred by searching for sequence similarity in databases and repositories of proteomic data [43, 44]. In a few cases, biochemical assays have been undertaken to unveil additional details regarding the functional role of newly evolved genes. For example, the protein encoded by the new *Drosophila* alcohol dehydrogenase gene *jingwei* shows modified substrate specificity in relation to different types of alcohols in comparison to the protein encoded by its parental gene *Adh* [45]. By combining enzymatic assays with site-directed mutagenesis, it has also been shown that a new pancreatic ribonuclease in colobine primates evolved to function in more acidic digestive environments [46].

Although the bioinformatic and biochemical approaches described above provide an initial characterization of a new gene's function, they do not resolve two key, intertwined aspects of its evolution: (i) how the gene becomes incorporated into pre-existing genetic networks; and (ii) how relevant the gene is for the organism's fitness. Similarly, patterns of protein sequence evolution compatible with the action of positive or negative selection (Box 2) are a proxy for functional relevance, but they do not precisely answer the above questions. Furthermore, merely demonstrating that a new gene has a different expression pattern than its parental gene is not sufficient to conclude that it has gained a new function, since cryptic promoters and signals for splicing and polyadenylation are known to be ubiquitous in intergenic regions [47, 48] and many cases of ectopic gene expression have been documented [49]. This is especially relevant for de novo genes, which are currently the subject of intense study across diverse phyla [23, 26, 33, 50]. To overcome these limitations, experiments in model organisms are essential, because they allow one to determine a new gene's function by examining phenotypes that result from perturbing the activity of the gene.

Box 2

Measuring the rate of protein evolution

Because of the redundancy of the genetic code, point mutations occurring within a protein-encoding sequence can be classified as either nonsynonymous or synonymous. The former alter the amino acid sequence of the encoded protein, while the latter do not.

Synonymous mutations are assumed to be selectively neutral and accumulate over time, resulting in sequence divergence between species. Nonsynonymous mutations, in contrast, are subject to natural selection and will only contribute to sequence divergence between species if they do not have a harmful effect on the organism. Otherwise, they will be eliminated from the population by negative (or ''purifying'') selection. After correcting for the number of sites in the sequence where either a nonsynonymous or a synonymous mutation could occur, the ratio of nonsynonymous to synonymous divergence between species (*Ka*/*Ks*) can be used as a measure of the rate of protein evolution [74]. Genes encoding functionally important proteins are expected to have *Ka*/*Ks* values close to zero, because most changes to their amino acid sequence are deleterious. Following a duplication event, new genes typically show elevated *Ka*/*Ks* relative to their parental genes. This could reflect a relaxation of functional constraint and/or positive selection favoring amino acid changes that lead to a new, beneficial function. In the latter case, functional constraint should increase after the new function has been optimized. Thus, *Ka*/*Ks* will be lower among alleles of the gene in the present day population than between copies of the gene from different species.

Can a newly evolved gene affect an organism's viability or fertility?

RNA interference (RNAi) is one method that can be used to determine whether or not new genes are required for successful development to the adult stage. This approach was used by Chen et al. [7] to inhibit the expression of young genes (those originating within the past 55 million years) in the *D. melanogaster* genome. In total, they examined 195 newly evolved genes and a random sample of 245 old genes. The proportion of essential genes in both age groups was approximately one third (Fig. 2), indicating that once young genes have been successfully incorporated into the genome, they can play roles as critical for the organism as old genes. The authors found that the fraction of essential genes is nearly constant across different age groups (Fig. 2) and concluded that even very recently generated genes can become essential. Further, examination of the temporal expression pattern of essential young genes revealed that most of them were expressed predominantly during the larval and pupal stages, which suggests that they are developmentally regulated. The comparison of essentiality between parental and derived genes showed that young essential genes can be generated with equal probability from essential or non-essential parental genes, and likewise, both essential and non-essential parental genes can give rise to essential or non-essential young genes (Fig. 2). The analysis of protein-protein interaction data supports an evolutionary dynamic in which newly evolved genes become integrated into gene networks, sometimes occupying key network positions that are involved in many interactions. Overall, these results emphasize that new essentiality can evolve very quickly, and therefore, genes associated with it represent an unparalleled reservoir of genetic variation that can facilitate species adaptation and divergence.

To date, no large-scale silencing experiments have been performed focusing on fertility. However, several case studies illustrate how newly evolved genes might impact reproduction. Two studies focused on genes unique to closely related species of the *D. melanogaster* species subgroup. The new gene *nsr* is involved in the regulation of mRNA processing of three adjacently located *Y*-linked genes that encode dynein heavy chains of the sperm axoneme. Knockout experiments revealed multiple sperm deficiencies when this gene is inactive. Interestingly, the parental gene of *nsr* is important for female fertility [51]. On the other hand, the gene *K81* is one of the few genes reported to have a paternal effect in *Drosophila* males. Males with mutations in this gene produce sperm able to fertilize oocytes, however the paternal chromosomes fail to segregate appropriately during the first zygotic division [52]. A third study in mice (genus *Mus*) reported the emergence of a non-coding RNA (ncRNA) gene, *Poldi*, which, when knocked out, results in males with testis of reduced weight and sperm of limited motility [50]. In these examples, the knock out of a newly evolved gene precludes a successful fertilization event. A fourth case study shows that new genes can even prevent fertilization by altering mating behavior. For example, knocking out the ncRNA gene *Sphinx* led to increased same-sex courtship in male *Drosophila* [53]. Based on its expression in different brain areas and genome-wide microarray experiments, *Sphinx* is suggested to participate in olfactory neuron mediated regulation of male courtship [54].

Controlled perturbation of gene function is not exempt from limitations. First, the resulting phenotype might be difficult to detect under laboratory conditions. This has been shown to be the case for a sizable fraction of genes in mice [55]. In fact, when further experiments were performed with one of

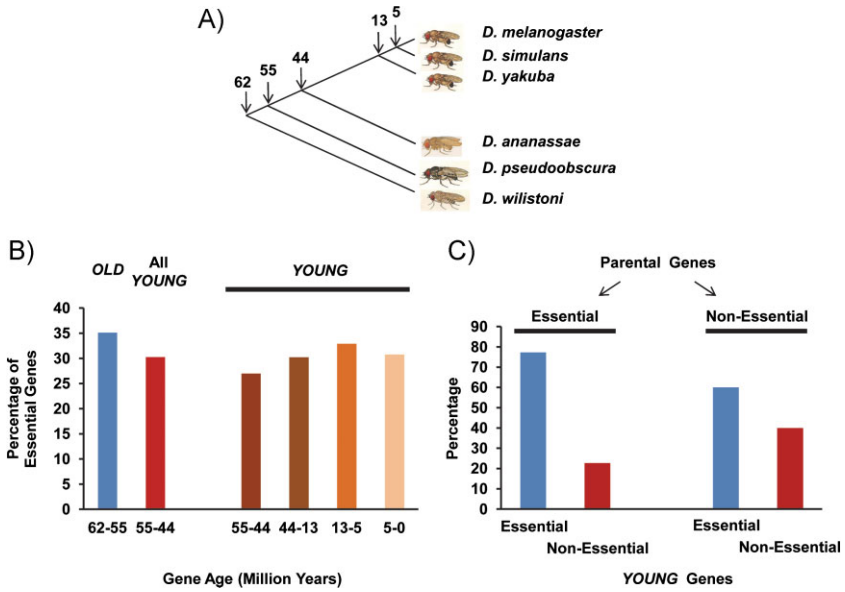


Figure 2. A: Phylogeny of the *Drosophila* species used to date the evolutionary age of the genes analyzed in Chen et al. [7]. The numbers at the nodes indicate the divergence times of the different lineages (in millions of years) [75]. **B**: Gene age is not correlated with essentiality in fruit flies. The fraction of essential genes is constant among different age groups across the genus *Drosophila*. **C**: The ratio of essential to non-essential young genes does not depend on whether or not the parental gene is essential. *Drosophila* images are from FlyBase (www.flybase.org).

the newly evolved *Drosophila* genes deemed as non-essential by Chen et al. [7], it was found that the gene plays a crucial role in determining life expectancy and fertility [56]. Another limitation is that perturbation of gene expression is only feasible in model organisms. In the case of humans, disease association studies can help to overcome this limitation. For example, the genes *DDX11* and *DPY19L2*, which are exclusive to chimps and humans [6], have been reported to underlie the Warsaw breakage syndrome [57] and male infertility problems [58], respectively.

Are chimeric genes special?

Although new genes may be created by various mechanisms (Box 1), the probability of gaining a new, beneficial function might not be equal for all kinds of newly evolved genes. Chimeric genes appear to be especially well-suited for the adoption of new functions and integration into pre-existing genetic networks. A property that chimeric genes possess, but regular duplicates lack, is the potential to create new combinations of polypeptide sequences at the time of their initial creation. This allows them to explore the possible protein sequence space much more efficiently than duplicate genes, which must rely on amino acid substitutions alone [59, 60]. Recently created chimeric proteins can have the combined functional attributes of their parental genes [61, 62], fueling phenotypic diversity, and therefore providing variation upon which natural selection may act [62, 63]. A possible mechanistic explanation is that chimeric proteins can remodel the protein-protein interaction network. In fact, protein-protein interactions can pose severe constraints that prevent evolutionary change because of pleoitropic effects. Newly evolved chimeric genes might offer an efficient, alternative path to relax this constraint, as has been proposed for the yeast mating pathway [64]. Peisajovich et al. [64] engineered 66 recombinant proteins by combining whole-domains from the 11 constituent genes of the pathway. Similarly, they engineered full gene duplications, domain duplications, and combinations of two unlinked but co-expressed domains. All of these recombinant peptides were transformed into a yeast strain that possessed the original gene set and the magnitude of the mating response monitored upon inducing its activation. Only chimeric proteins contributed to the diversification of the signaling phenotype, sometimes resulting in a higher mating efficiency than that of the wild-type strain. This advantage was in some cases large enough to offset detrimental effects in other phenotypes, e.g. growth rate, which is related to the high osmolarity response pathway. This observation stresses the pleiotropic nature associated with chimeric proteins.

Despite their immediate provision of raw material for natural selection to act upon and their enhanced effectiveness in exploring sequence space, chimeric genes are not always beneficial. Often, chimeric genes have detrimental effects on their carriers because they encode dysfunctional proteins that alter fundamental biological process such as cell cycle regulation [65, 66]. It has been estimated that only 1.4% of the chimeric genes that arose in the *D. melanogaster* lineage have been preserved [62]. Nevertheless, gene fusion represents a common evolutionary mechanism of new gene creation across diverse phyla including plants [67], primates [68], fish [69], nematodes [70], and insects [25, 62]. Interestingly, genes particularly important for the ecology of the species have been shown to be involved in recurrent and phylogenetically independent events of chimeric gene formation. An example of this is provided by the *Drosophila* alcohol dehydrogenase gene *Adh*, which functions in the catabolism of alcohols and has been involved in at least four independent gene fusion events [15, 71–73].

In addition to combining protein-encoding sequences from two different sources, many chimeric genes also contain new or altered regulatory information. For example, the protein encoded by a chimeric gene of the RGP family in humans

shows discrete localization in the cytoplasm, while the protein encoded by the parental gene is found only in the nuclear envelope [20]. Analyzing a conservative set of chimeric genes present in the genome of *D. melanogaster*, it has been found that the formation of chimeric genes also enables new combinations of regulatory and RNA stability motifs, as well as cellular targeting signals, which can facilitate the evolution of novel expression profiles and alter the cellular context of the encoded peptide [63]. Importantly, half of the chimeric genes analyzed were found to involve mid-protein domain breaks instead of whole-domain breaks, which indicates that the scale of effective modularity is much smaller than previously thought.

Conclusions

Vast amounts of DNA sequence data and the development of in silico tools to analyze and compare genomes have enabled the identification of new genes that are unique to a species or a phylogenetic lineage. To date, case studies have focused mostly on the structure, pattern of sequence evolution, and preliminary characterization of the function of newly evolved genes. Whether or not these functions are relevant for the species and how they impact the organism's fitness are only starting to be explored. Perturbations of gene function, associations with aberrant phenotypes, and more detailed studies of how the encoded proteins are integrated into protein-protein interaction networks will play a fundamental role in determining the functional relevance of new genes and their contribution to the adaptation and divergence of species. Genes that combine diverse functional attributes of pre-existing genes at the time of their origination (e.g. chimeric genes) appear to be key players in these processes and represent an exciting area in genome and evolutionary research.

Acknowledgments

We thank Kania Gandasetiawan, Carolus Chan, and two reviewers for useful comments on the manuscript. This work was supported by NSF grant (DEB-0949365) to J.R.

References

1. **Nakabachi A**, **Yamashita A**, **Toh H**, **Ishikawa H**, et al. 2006. The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* **314**: 267.
2. **Yu J**, **Hu S**, **Wang J**, **Wong GK**, et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**: 79–92.
3. **Goff SA**, **Ricke D**, **Lan TH**, **Presting G**, et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**: 92–100.
4. **Kumar S**, **Filipski A**, **Swarna V**, **Walker A**, et al. 2005. Placing confidence limits on the molecular age of the human-chimpanzee divergence. *Proc Natl Acad Sci USA* **102**: 18842–7.
5. **Demuth JP**, **De Bie T**, **Stajich JE**, **Cristianini N**, et al. 2006. The evolution of mammalian gene families. *PLoS One* **1**: e85.
6. **Zhang YE**, **Landback P**, **Vibranovski MD**, **Long M.** 2011. Accelerated recruitment of new brain development genes into the human genome. *PLoS Biol* **9**: e1001179.
7. **Chen S**, **Zhang YE**, **Long M.** 2010. New genes in *Drosophila* quickly become essential. *Science* **330**: 1682–5.
8. **Tautz D**, **Domazet-Loso T.** 2011. The evolutionary origin of orphan genes. *Nat Rev Genet* **12**: 692–702.
9. **Rakic P.** 2009. Evolution of the neocortex: a perspective from developmental biology. *Nat Rev Neurosci* **10**: 724–35.
10. **Kaessmann H.** 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res* **20**: 1313–26.
11. **Hahn MW**, **Han MV**, **Han SG.** 2007. Gene family evolution across 12 *Drosophila* genomes. *PLoS Genet* **3**: e197.
12. **Schrider DR**, **Hahn MW.** 2010. Gene copy-number polymorphism in nature. *Proc Biol Sci* **277**: 3213–21.
13. **Brosius J.** 1991. Retroposons—seeds of evolution. *Science* **251**: 753.
14. **Betran E**, **Thornton K**, **Long M.** 2002. Retroposed new genes out of the X in *Drosophila*. *Genome Res* **12**: 1854–9.
15. **Long M**, **Langley CH.** 1993. Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. *Science* **260**: 91–5.
16. **Nurminsky DI**, **Nurminskaya MV**, **De Aguiar D**, **Hartl DL.** 1998. Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. *Nature* **396**: 572–5.
17. **Rogers RL**, **Bedford T**, **Lyons AM**, **Hartl DL.** 2010. Adaptive impact of the chimeric gene *Quetzalcoatl* in *Drosophila melanogaster*. *Proc Natl Acad Sci USA* **107**: 10943–8.
18. **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.
19. **Courseaux A**, **Nahon JL.** 2001. Birth of two chimeric genes in the Hominidae lineage. *Science* **291**: 1293–7.
20. **Ciccarelli FD**, **von Mering C**, **Suyama M**, **Harrington ED**, et al. 2005. Complex genomic rearrangements lead to novel primate gene function. *Genome Res* **15**: 343–51.
21. **Brennan G**, **Kozyrev Y**, **Hu SL.** 2008. TRIMCyp expression in Old World primates *Macaca nemestrina* and *Macaca fascicularis*. *Proc Natl Acad Sci USA* **105**: 3569–74.
22. **Liu SL**, **Zhuang Y**, **Zhang P**, **Adams KL.** 2009. Comparative analysis of structural diversity and sequence evolution in plant mitochondrial genes transferred to the nucleus. *Mol Biol Evol* **26**: 875–91.
23. **Levine MT**, **Jones CD**, **Kern AD**, **Lindfors HA**, 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 USA* **103**: 9935–9.
24. **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**: 1131–7.
25. **Zhou Q**, **Zhang G**, **Zhang Y**, **Xu S**, et al. 2008. On the origin of new genes in *Drosophila*. *Genome Res* **18**: 1446–55.
26. **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.
27. **Knowles DG**, **McLysaght A.** 2009. Recent de novo origin of human protein-coding genes. *Genome Res* **19**: 1752–9.
28. **Toll-Riera M**, **Bosch N**, **Bellora N**, **Castelo R**, et al. 2009. Origin of primate orphan genes: a comparative genomics approach. *Mol Biol Evol* **26**: 603–12.
29. **Xiao W**, **Liu H**, **Li Y**, **Li X**, et al. 2009. A rice gene of de novo origin negatively regulates pathogen-induced defense response. *PLoS One* **4**: e4603.
30. **Li D**, **Dong Y**, **Jiang Y**, **Jiang H**, et al. 2010. 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.
31. **Yang Z**, **Huang J.** 2011. De novo origin of new genes with introns in *Plasmodium vivax*. *FEBS Lett* **585**: 641–4.
32. **Li CY**, **Zhang Y**, **Wang Z**, **Zhang Y**, et al. 2010. A human-specific de novo protein-coding gene associated with human brain functions. *PLoS Comput Biol* **6**: e1000734.
33. **Wu DD**, **Irwin DM**, **Zhang YP.** 2011. De novo origin of human protein-coding genes. *PLoS Genet* **7**: e1002379.
34. **Petrov DA**, **Lozovskaya ER**, **Hartl DL.** 1996. High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384**: 346–9.
35. **McDonald JH**, **Kreitman M.** 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**: 652–4.
36. **Hahn MW.** 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. *J Hered* **100**: 605–17.
37. **Bai Y**, **Casola C**, **Betran E.** 2008. Evolutionary origin of regulatory regions of retrogenes in *Drosophila*. *BMC Genomics* **9**: 241.
38. **Parkinson H**, **Sarkans U**, **Kolesnikov N**, **Abeygunawardena N**, et al. 2010. ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res* **39**: D1002–4.

39. **Barrett T**, **Troup DB**, **Wilhite SE**, **Ledoux P**, et al. 2010. NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res* **39**: D1005–10.
40. **Chintapalli VR**, **Wang J**, **Dow JA.** 2007. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. *Nat Genet* **39**: 715–20.
41. **Dash S**, **Van Hemert J**, **Hong L**, **Wise RP**, et al. 2012. PLEXdb: gene expression resources for plants and plant pathogens. *Nucleic Acids Res* **40**: D1194–201.
42. **Finger JH**, **Smith CM**, **Hayamizu TF**, **McCright IJ**, et al. 2011. The mouse Gene Expression Database (GXD): 2011 update. *Nucleic Acids Res* **39**: D835–41.
43. **Deutsch EW**, **Lam H**, **Aebersold R.** 2008. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep* **9**: 429–34.
44. **Vizcaino JA**, **Cote R**, **Reisinger F**, **Foster JM**, et al. 2009. A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics* **9**: 4276–83.
45. **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.
46. **Zhang J.** 2006. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nat Genet* **38**: 819–23.
47. **van Bakel H**, **Hughes TR.** 2009. Establishing legitimacy and function in the new transcriptome. *Brief Funct Genomic Proteomic* **8**: 424–36.
48. **van Bakel H**, **Nislow C**, **Blencowe BJ**, **Hughes TR.** 2010. Most ''dark matter'' transcripts are associated with known genes. *PLoS Biol* **8**: e1000371.
49. **Clark MB**, **Amaral PP**, **Schlesinger FJ**, **Dinger ME**, et al. 2011. The reality of pervasive transcription. *PLoS Biol* **9**: e1000625 (discussion e1102).
50. **Heinen TJ**, **Staubach F**, **Haming D**, **Tautz D.** 2009. Emergence of a new gene from an intergenic region. *Curr Biol* **19**: 1527–31.
51. **Ding Y**, **Zhao L**, **Yang S**, **Jiang 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.
52. **Loppin B**, **Lepetit D**, **Dorus S**, **Couble P**, et al. 2005. Origin and neofunctionalization of a *Drosophila* paternal effect gene essential for zygote viability. *Curr Biol* **15**: 87–93.
53. **Dai H**, **Chen Y**, **Chen S**, **Mao Q**, 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.
54. **Chen Y**, **Dai H**, **Chen S**, **Zhang L**, et al. 2011. Highly tissue specific expression of *Sphinx* supports its male courtship related role in *Drosophila melanogaster*. *PLoS One* **6**: e18853.
55. **Bult CJ**, **Kadin JA**, **Richardson JE**, **Blake JA**, et al. 2010. The Mouse Genome Database: enhancements and updates. *Nucleic Acids Res* **38**: D586–92.
56. **Chen S**, **Yang H**, **Krinsky BH**, **Zhang A**, et al. 2011. Roles of young serine-endopeptidase genes in survival and reproduction revealed rapid evolution of phenotypic effects at adult stages. *Fly (Austin)* **5**: 345–51.
57. **van der Lelij P**, **Chrzanowska KH**, **Godthelp BC**, **Rooimans MA**, et al. 2010. Warsaw breakage syndrome, a cohesinopathy associated with mutations in the XPD helicase family member DDX11/ChlR1. *Am J Hum Genet* **86**: 262–6.
58. **Harbuz R**, **Zouari R**, **Pierre V**, **Ben Khelifa M**, et al. 2011. A recurrent deletion of DPY19L2 causes infertility in man by blocking sperm head elongation and acrosome formation. *Am J Hum Genet* **88**: 351–61.
59. **Cui Y**, **Wong WH**, **Bornberg-Bauer E**, **Chan HS.** 2002. Recombinatoric exploration of novel folded structures: a heteropolymer-based model of protein evolutionary landscapes. *Proc Natl Acad Sci USA* **99**: 809–14.
60. **Carneiro M**, **Hartl DL.** 2010. Colloquium papers: adaptive landscapes and protein evolution. *Proc Natl Acad Sci USA* **107**: 1747–51.
61. **Patthy L.** 2003. Modular assembly of genes and the evolution of new functions. *Genetica* **118**: 217–31.
62. **Rogers RL**, **Bedford T**, **Hartl DL.** 2009. Formation and longevity of chimeric and duplicate genes in *Drosophila melanogaster*. *Genetics* **181**: 313–22.
63. **Rogers RL**, **Hartl DL.** 2012. Chimeric genes as a source of rapid evolution in *Drosophila melanogaster*. *Mol Biol Evol* **29**: 517–29.
64. **Peisajovich SG**, **Garbarino JE**, **Wei P**, **Lim WA.** 2010. Rapid diversification of cell signaling phenotypes by modular domain recombination. *Science* **328**: 368–72.
65. **Mitelman F**, **Johansson B**, **Mertens F.** 2007. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer* **7**: 233–45.
66. **Maher CA**, **Kumar-Sinha C**, **Cao X**, **Kalyana-Sundaram S**, et al. 2009. Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**: 97–101.
67. **Wang W**, **Zheng H**, **Fan C**, **Li J**, et al. 2006. High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell* **18**: 1791–802.
68. **Marques-Bonet T**, **Girirajan S**, **Eichler EE.** 2009. The origins and impact of primate segmental duplications. *Trends Genet* **25**: 443–54.
69. **Fu B**, **Chen M**, **Zou M**, **Long M**, et al. 2010. The rapid generation of chimerical genes expanding protein diversity in zebrafish. *BMC Genomics* **11**: 657.
70. **Katju V**, **Lynch M.** 2006. On the formation of novel genes by duplication in the *Caenorhabditis elegans* genome. *Mol Biol Evol* **23**: 1056–67.
71. **Jones CD**, **Begun DJ.** 2005. Parallel evolution of chimeric fusion genes. *Proc Natl Acad Sci USA* **102**: 11373–8.
72. **Jones CD**, **Custer AW**, **Begun DJ.** 2005. Origin and evolution of a chimeric fusion gene in *Drosophila subobscura, D. madeirensis* and *D. guanche*. *Genetics* **170**: 207–19.
73. **Shih HJ**, **Jones CD.** 2008. Patterns of amino acid evolution in the *Drosophila ananassae* chimeric gene, siren, parallel those of other *Adh*-derived chimeras. *Genetics* **180**: 1261–3.
74. **Hurst LD.** 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet* **18**: 486.
75. **Tamura K**, **Subramanian S**, **Kumar S.** 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol Biol Evol* **21**: 36–44.