What Signatures Dominantly Associate with Gene Age?

Hongyan Yin[1,2,3,†], Guangyu Wang[1,2,3,†], Lina Ma[1,2], Soojin V. Yi[4], and Zhang Zhang[1,2,3,*]

[1]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, Beijing, China

[2]BIG Data Center, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, Beijing, China

[3]University of Chinese Academy of Sciences, Beijing, China

[4]School of Biology, Georgia Institute of Technology, Atlanta

*Corresponding author: E-mail: zhangzhang@big.ac.cn.

[†]These authors contributed equally to this work.

Accepted: August 31, 2016

Abstract

As genes originate at different evolutionary times, they harbor distinctive genomic signatures of evolutionary ages. Although previous studies have investigated different gene age-related signatures, what signatures dominantly associate with gene age remains unresolved. Here we address this question via a combined approach of comprehensive assignment of gene ages, gene family identification, and multivariate analyses. We first provide a comprehensive and improved gene age assignment by combining homolog clustering with phylogeny inference and categorize human genes into 26 age classes spanning the whole tree of life. We then explore the dominant age-related signatures based on a collection of 10 potential signatures (including gene composition, gene length, selection pressure, expression level, connectivity in protein–protein interaction network and DNA methylation). Our results show that GC content and connectivity in protein–protein interaction network (PPIN) associate dominantly with gene age. Furthermore, we investigate the heterogeneity of dominant signatures in duplicates and singletons. We find that GC content is a consistent primary factor of gene age in duplicates and singletons, whereas PPIN is more strongly associated with gene age in singletons than in duplicates. Taken together, GC content and PPIN are two dominant signatures in close association with gene age, exhibiting heterogeneity in duplicates and singletons and presumably reflecting complex differential interplays between natural selection and mutation.

Key words: gene age, signature, GC content, PPIN, principle component analysis.

Introduction

Birth of new genes is associated with events of gene duplication (Betran et al. 2002; Long et al. 2003), horizontal gene transfer (Keeling and Palmer 2008), extant gene fragments (Gilbert 1978), and de novo creations from noncoding DNA/RNA (Knowles and McLysaght 2009; Toll-Riera et al. 2009). It is considered that the birth of new genes is one of several primary mechanisms underlying the evolution of novel functions in biological systems, and often facilitating adaptive evolution (Kaessmann 2010). Accordingly, genes, that are birthed and fixed into a species at specific evolutionary time, are left with distinctive age-related signature (ARS) in the genome. Therefore, deciphering ARS in molecular sequences holds great significance in better understanding molecular evolutionary processes and unveiling the underlying mechanisms that drive young genes to become indispensible integrants coupled with novel phenotypes and biological diversities (Long et al. 2013).

To date, attempts have been made to address this issue by detecting diverse ARS. These studies revealed that young genes are shorter, have fewer introns (Wolf et al. 2009), relate closely with the birth of new binding sites (Ni et al. 2012) and harbor more premature termination codon mutations (Yang et al. 2015). Furthermore, young genes possess fewer interactions with other genes (Zhang et al. 2015) and tend to play less essential functional roles compared with old genes (Chen et al. 2012). Additionally, young human genes are likely to present distinct temporal and spatial expression patterns (Long et al. 2013; Popadin et al. 2014). It is reported that young genes evolve more rapidly (Alba and Castresana 2005; Wolf et al. 2009) and experience more variable selection pressure than old genes (Vishnoi et al. 2010). Moreover, a

recent study has further shown that young and old duplicates differ strikingly in their DNA methylation (Keller and Yi 2014).

Although different aged genes differ in multiple ARS as mentioned earlier, the relative significance of different ARS is not known and it remains unresolved what signatures dominantly associate with gene age. Additionally, previous studies on age identification mainly employed similarity search that have been reported to be error-prone (Alba and Castresana 2007; Tautz and Domazet-Loso 2011; Moyers and Zhang 2015, 2016) and determined gene age based on a rough evolutionary time-scale (Domazet-Loso and Tautz 2008; Wolf et al. 2009; Zhang et al. 2010). As a result, discriminating origins of divergent homologs and capturing important evolutionary events have been difficult. In this study, we provide a newly generated, comprehensive and improved gene age identification by combining homolog clustering with phylogeny inference. Accordingly, we determine gene age at an extremely refined evolutionary time-scale and categorize human genes into 26 evolutionary age classes spanning the whole tree of life. Using this age identification, we explore dominant ARS in the human genome based on a collection of 10 potential ARS and further investigate the heterogeneity of dominant ARS in duplicates and singletons.

Results and Discussion

Age Identification of Human Genes

Improving upon previous studies on age identification (Domazet-Loso and Tautz 2008; Wolf et al. 2009; Zhang et al. 2010), here we combine homolog clustering with phylogeny inference to identify gene ages (see "Methods" section) and categorize human genes into 26 age classes ranging from archaea/bacteria (age class 26) to human (age class 1), spanning an extremely long evolutionary time-scale of ~4,000 million years (supplementary tables S1 and S2 and fig. S1, Supplementary Material online). Our classification provides the most refined evolutionary gene-age classes so far, compared with previous studies where genes were classified into seven classes in (Wolf et al. 2009), 11 classes in (Cai and Petrov 2010) and 19 classes in (Domazet-Loso and Tautz 2008). Although our refined age classification, by virtue of increased number of classes, could misclassify ages of some genes whose sequence and annotation in the current genome assemblies include errors, our results on gene age identification present three major improvements. First, previous studies for gene age identification were typically based on homolog similarity, and thus not well suited to effectively differentiate the origins of paralogs. In comparison, our study, by utilizing homolog clustering with phylogeny inference, is able to confidently identify evolutionary ages of paralogous genes.

Second, we utilize an extremely refined phylogenetic framework consisting of 26 age classes, encompassing major evolutionary events from unicellular organisms to human. Consequently, it is capable to investigate gene loss events (a gene loss event is determined and counted when a gene is present at certain evolutionary time, but absent afterwards) in a more detailed manner based on our age identification results. For instance, a previous study has reported that genes are lost after the divergence of human and rodents (Blomme et al. 2006). Contrastingly, our results show that those specific genes are heavily lost at the origination time of primates and scandentia (supplementary fig. S2, Supplementary Material online), yielding a higher resolution determination of important evolutionary events. Meanwhile, among the 26 age classes, primate-specific evolutionary time-scale is well separated into seven different age classes (namely tarsiiformes, platyrrhini, cercopithecidae, hylobatidaee, pongo, gorillae and human), which is of great significance for better understanding details of primate evolutionary processes and innovation of primate-specific genes. For instance, MYEOV (ENSG0000017292), a gene that has been reported to de novo arise from noncoding RNA in human-specific lineage (Xie et al. 2012), actually arose at age class 4, namely, hominoid-specific lineage, indicating that its transition from noncoding RNA to a coding gene is, more precisely, occurred at the origin of hominoidea.

Third, our method based on a phylogenetic framework features effective inference of evolutionary time of gene duplication events, allowing confident age assignments of paralogs. Accordingly, we find that 11% duplication events can be traced back to the origin of metazoan ~900 million years ago (Mya) (supplementary fig. S3, Supplementary Material online) and 16% duplication events are assigned to the origin of vertebrate ~450 Mya (supplementary fig. S3, Supplementary Material online), indicating that the origins of multicellularity and vertebrate are fundamental, presumably with key innovations for the emergence of human genes. These results are in good accordance with the hypothesis that hierarchical complexity increases at the origin of multicellularity (Rainey 2007) and that duplication events, including whole genome duplication, are major evolutionary forces underlying vertebrate genome evolution (Blomme et al. 2006).

Dominant ARS in Human Gene

As mentioned earlier, birth of genes leads to different ARS at multiple omics levels, including genomics, transcriptomics, epigenetics, etc. Based on our age identification, here we incorporate a total of 10 potential ARS, including nucleotide composition (GC/AG content) of entire CDS (coding sequence), sequence length, CUB, expression level, natural selection inferred from nonsynonymous/synonymous substitution ratio (Ka/Ks), DNA methylation, and PPIN (supplementary fig. S4, Supplementary Material online). Since these signatures are highly interdependent (Kim and Yi 2007; Park et al. 2012), correlation analysis cannot be used to identify dominant

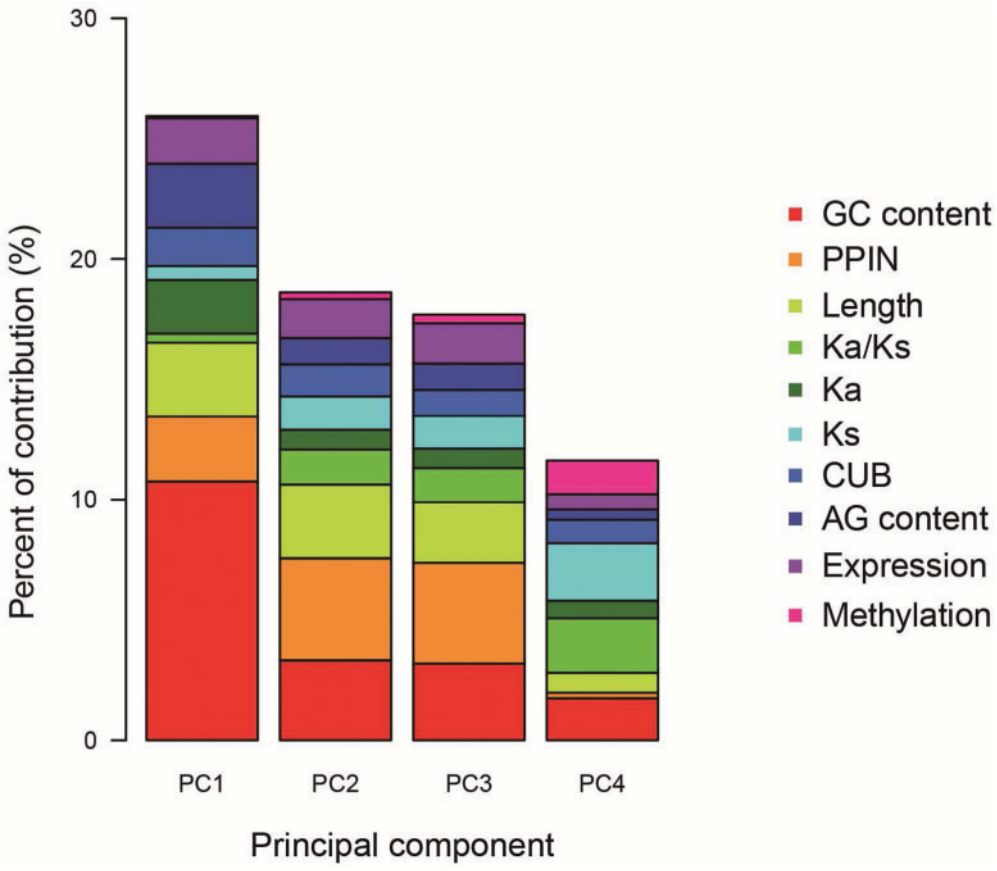

Fig. 1.—Principal component analyses on gene age. The corresponding numerical results were summarized into supplementary table S4, Supplementary Material online.

signatures associated closely with gene age (supplementary table S3, Supplementary Material online). Therefore, we perform principal component analysis (PCA), a widely used statistical method that is able to transform a set of possibly correlated variables into a set of linearly uncorrelated principal components, to decipher which signatures are highly dominant with gene age. According to the PCA results, the first four principal components account for 74% of the variance and the first two principal components are able to explain ~44.55% of the variance (fig. 1A and supplementary table S4, Supplementary Material online). Notably, the first component is mainly dominated by GC content (41.46%) and the second component is largely determined by PPIN (22.79%). Our results clearly show that, albeit ARS are evolutionarily confounded and interrelated, GC content and PPIN are two dominant signatures associating closely with gene age. Additionally, to avoid bias due to the large number of genes in the strata of unicellular organisms (supplementary fig. S5, Supplementary Material online), we re-sample the same percentage of genes from unicellular organisms as that of multicellular organisms (supplementary table S5, Supplementary Material online) and consistently obtain the similar results that GC content and PPIN are dominant ARS.

Duplication is a main driving mechanism in the birth of new genes (Gu et al. 2002; Zhang 2003). It is reported that singletons evolve more rapidly (Jordan et al. 2004) and tend to have more consistent expression profiles than duplicates (Li et al. 2005). Given these observations, we hypothesize that dominant ARS may be different between duplicates and singletons. Therefore, we further perform PCA separately on singletons and duplicates. Our results show that in singletons (fig. 2A and supplementary table S6, Supplementary Material online), consistent with previous results, the first component is determined mainly by GC content (35.21%) and the second component is determined mainly by PPIN (26.14%). Intriguingly, in duplicates (fig. 2B and supplementary table S7, Supplementary Material online), the first component is determined mainly by GC content (35.79%) and the second component is still determined by GC content (30.63%). Together, GC content consistently dominates as a primary signature with gene age in duplicates and singletons, whereas PPIN dominates with gene age more significantly in singletons than in duplicates. These results indicate that duplicates and

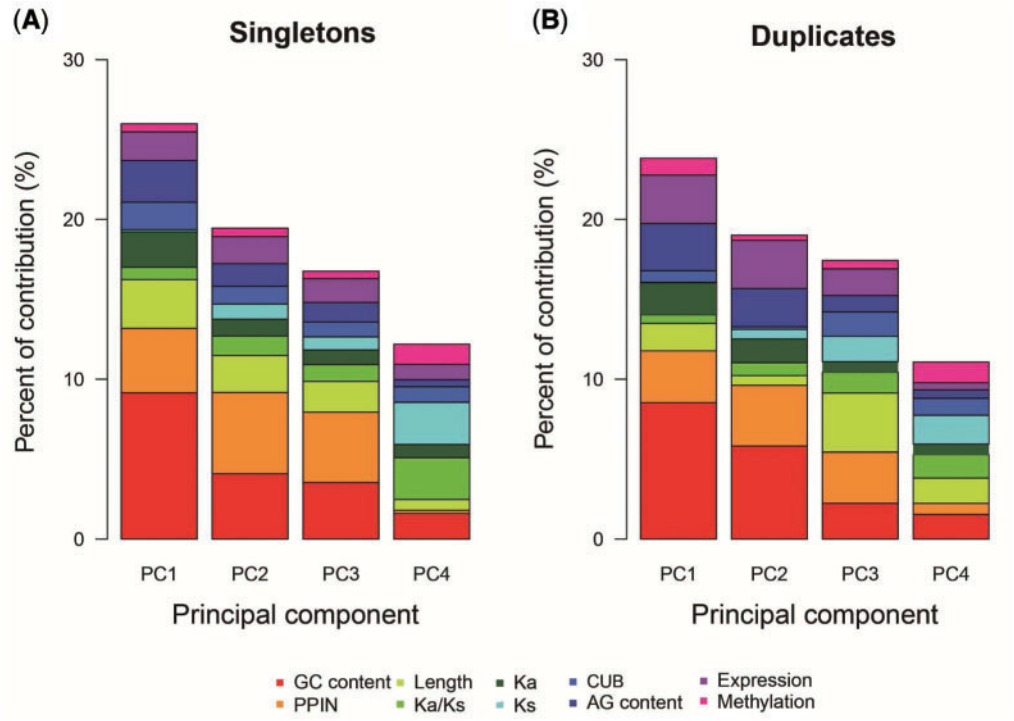

Fig. 2.—Principal component analyses on gene age in singletons and duplicates. The corresponding numerical results were summarized into supplementary tables S6 and S7, Supplementary Material online.

singletons may experience diverse evolutionary forces and yield different dominant signatures of gene age.

It is well documented that GC content, as one of the most fundamental gene features, is highly correlated with multiple factors [including mutation (Fryxell and Moon 2005), selection for specific synonymous codons for translation efficiency and accuracy (Plotkin and Kudla 2011), horizontal gene transfer (Philippe and Douady 2003), methylation modification (Bird 1986; Elango et al. 2008), and gene density (Duret et al. 1995), etc.]. It is notable that even though we separately examined factors known to associate with GC content, including gene length, expression, codon usage, selection strength and methylation, we still observe the dominant significance of GC content in the evolution. Therefore, our results indicate significant effects of GC content apart from the aforementioned factors. For example, replication dynamics correlates with GC content (Kenigsberg et al. 2016) and GC-biased gene conversion specifically in highly recombining genomic regions affects the genomes of most bacterial species (Lassalle et al. 2015) and many eukaryotes (Webster and Hurst 2012), further providing the possibility that the dominance of GC content helps shape the genome characterization universally. Consistently and strikingly, our results demonstrate that regardless of being duplicates or singletons, GC content is an overwhelmingly dominant signature associating closely with gene age. Conforming to this point, additional evidence has shown by a recent study that de novo new genes originating from long noncoding RNAs present heterogeneity in GC content (Chen et al. 2015).

It has been reported that more than one-third of known regulatory interactions in yeast (Teichmann and Babu 2004) and average 27% interaction networks for primate-specific young genes in human (Zhang et al. 2015) are inherited from their parental genes after duplication, so that duplication is a significant contributor of gene interaction network (Middendorf et al. 2005). In contrast to duplicates that have inherited PPIN from parental copies, singletons have little interactions at their early evolutionary stage, but, over time, they are gradually integrated into gene interaction networks to acquire biological functions. As genes evolve and age in the genome, therefore, singletons may experience more dramatic variations in PPIN than duplicates. Indeed, singletons do exhibit a much larger variability in PPIN compared with duplicates (supplementary fig. S6, Supplementary Material online). Consequently, even though old genes (including duplicates and singletons) tend to be highly connected in PPIN (Zhang et al. 2015), PPIN appears to be a more important signature of evolutionary age in singletons. Taken together, GC content and PPIN are two dominant signatures in close association with gene age, yet exhibiting heterogeneity in duplicates and singletons and presumably reflecting complex differential interplays between natural selection and mutation as they age.

Methods

Age Definition and Identification

For a given human gene, age was defined based on the presence of its ortholog in a wide range of species. We downloaded protein sequences from Ensembl (http://www.ensembl.org; supplementary table S1, Supplementary Material online) and obtained a collection of nonredundant proteins by only keeping longest splicing variants (supplementary table S1 and fig. S1, Supplementary Material online). BLAST searches were constructed for all nonredundant proteins (*E*-value $< 10^{-3}$). Furthermore, we conducted homolog clustering using the Markov Cluster algorithm (inflation value = 1.5) with OrthoMCL (Li et al. 2003) after loading BLAST results into MySQL database. Consequently, we assigned all resulting proteins into 35,948 homolog clusters. Among them, 12,493 singleton groups and 2,142 duplicate groups included human homologs. To infer the orthology relationships for duplicate group, multiple sequence alignments were conducted by MAFFT (Katoh and Standley 2013) and spurious sequences or poorly aligned regions were removed by trimAl (Capella-Gutierrez et al. 2009). Furthermore, we carried out phylogenetic inferences by phyML (Guindon et al. 2010) with bootstrap resampling tests by 100 times and utilized RIO (Resampled Inference of Orthologs; reliability values > 0.6) (Zmasek and Eddy 2002) for automated phylogeny inference to estimate the reliability of orthology assignments. As a consequence, we classified all human genes (including singletons and duplicates) into 25 age classes from the origin of eukaryotes, spanning ~1,500 Mya (age class 25) to human. Moreover, we used PANTHER (Mi et al. 2013) to determine orthology relationships between human genes and archaeal/bacterial genes. A human gene was assigned to age class 26 originating from ~4,000 Mya if its orthologs were detected in at least two archaeal/bacterial organisms (given the possibility of horizontal gene transfer). Detailed results of age identification for all human genes were tabulated in supplementary table S2, Supplementary Material online. In addition, gene ontology (GO) enrichment analyses were conducted and the corresponding results were summarized in supplementary table S8, Supplementary Material online.

Data Collection

We used homolog relationships between *Homo sapiens* and *Mus musculus* from NCBI HomoloGene database (http://www.ncbi.nlm.nih.gov/homologene) and obtained gene expression profiles across human 32 tissues from (Uhlen et al. 2015). We collected methylation data from GSE database with accession number GSE31848, including eight somatic cell samples (GSM868007-GSM868014) (Nazor et al. 2012). We retrieved PPIN data from (Cowley et al. 2012).

Estimation of Selection Pressure, Codon Usage Bias and Methylation Level

KaKs_Calculator (Zhang et al. 2006) was adopted to calculate nonsynonymous and synonymous substitution rates for human–mouse orthologs. Codon Deviation Coefficient (Zhang et al. 2012) was used to measure CUB for human genes as well as their orthologs among different species (supplementary fig. S7, Supplementary Material online). Methylation levels were estimated through an R package named Illumina Methylation Analyzer (Wang et al. 2012). For a given gene, DNA methylation level was averaged over its four regions including gene body region, promoter region, 5′- and 3′-UTR.

Principal Component Analysis

We used R for principle component analysis (package: pls). After logarithm transformation of four features including gene length, expression level, methylation level and PPIN, all features were scaled and normalized into [0, 1].

Supplementary Material

Supplementary figures S1–S7 and tables S1–S8 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by grants from The Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13040500 to Z.Z.), National Programs for High Technology Research and Development (863 Program; 2015AA020108 and 2012AA020409 to Z.Z.), International Partnership Program of the Chinese Academy of Sciences (153F11KYSB20160008), National Natural Science Foundation of China (Grant No. 31200978 to L.M.) and the "100-Talent Program" of Chinese Academy of Sciences (awarded to Z.Z.). This study was also supported by an NSF grant (SBE-131719), an NIH grant (1R01MH103517-01A1), and a Zoo Atlanta-Georgia Tech collaborative grant to S.V.Y.

Literature Cited

Alba MM, Castresana J. 2005. Inverse relationship between evolutionary rate and age of mammalian genes. Mol Biol Evol. 22:598–606.

Alba MM, Castresana J. 2007. On homology searches by protein Blast and the characterization of the age of genes. BMC Evol Biol. 7:53.

Betran E, Thornton K, Long M. 2002. Retroposed new genes out of the X in Drosophila. Genome Res. 12:1854–1859.

Bird AP. 1986. Cpg-rich islands and the function of DNA methylation. Nature 321:209–213.

Blomme T, et al. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol. 7:R43.

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.
Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.
Chen JY, et al. 2015. Emergence, retention and selection: a trilogy of origination for functional de novo proteins from ancestral LncRNAs in primates. Plos Genetics 11e:1005391.
Chen WH, Trachana K, Lercher MJ, Bork P. 2012. Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. Mol Biol Evol. 29:1703–1706.
Cowley MJ, et al. 2012. PINA v2.0: mining interactome modules. Nucleic Acids Res. 40:D862–D865.
Domazet-Loso T, Tautz D. 2008. An ancient evolutionary origin of genes associated with human genetic diseases. Mol Biol Evol. 25:2699–2707.
Duret L, Mouchiroud D, Gautier C. 1995. Statistical-analysis of vertebrate sequences reveals that long genes are scarce in Gc-rich isochores. J Mol Evol. 40:308–317.
Elango N, Kim SH, Vigoda E, Yi SV, Progra NCS. 2008. Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation. PLoS Comput Biol. 4:e1000015.
Fryxell KJ, Moon WJ. 2005. CpG mutation rates in the human genome are highly dependent on local GC content (vol 22, pg 650, 2005). Mol Biol Evol. 22:1159–1159.
Gilbert W. 1978. Why genes in pieces. Nature 271:501–501.
Gu X, Wang YF, Gu JY. 2002. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. Nat Genet. 31:205–209.
Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307–321.
Jordan IK, Wolf YI, Koonin EV. 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. BMC Evol Biol. 4:22.
Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. Genome Res. 20:1313–1326.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30:772–780.
Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet. 9:605–618.
Keller TE, Yi SV. 2014. DNA methylation and evolution of duplicate genes. Proc Natl Acad Sci U S A. 111:5932–5937.
Kenigsberg E, et al. 2016. The mutation spectrum in genomic late replication domains shapes mammalian GC content. Nucleic Acids Res. 44:4222–4232.
Kim SH, Yi SV. 2007. Understanding relationship between sequence and functional evolution in yeast proteins. Genetica 131:151–156.
Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. Genome Res. 19:1752–1759.
Lassalle F, Perian S, Bataillon T, Nesme X, Duret L, Daubin V. 2015. GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. Plos Genet. 11(2):e1004941.
Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.
Li WH, Yang J, Gu X. 2005. Expression divergence between duplicate genes. Trends Genet. 21:602–607.
Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. Nat Rev Genet. 4:865–875.
Long MY, VanKuren NW, Chen SD, Vibranovski MD. 2013. New gene evolution: little did we know. Annu Rev Genet. 47(47):307–333.
Mi HY, Muruganujan A, Thomas PD. 2013. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 41:D377–D386.
Middendorf M, Ziv E, Wiggins CH. 2005. Inferring network mechanisms: the *Drosophila melanogaster* protein interaction network. Proc Natl Acad Sci U S A. 102:3192–3197.
Moyers BA, Zhang JZ. 2015. Phylostratigraphic bias creates spurious patterns of genome evolution. Mol Biol Evol. 32:258–267.
Moyers BA, Zhang JZ. 2016. Evaluating phylostratigraphic evidence for widespread de novo gene birth in genome evolution. Mol Biol Evol. 33:1245–1256.
Nazor KL, et al. 2012. Recurrent variations in DNA methylation in human pluripotent stem cells and their differentiated derivatives. Cell Stem Cell 10:620–634.
Ni X, et al. 2012. Adaptive evolution and the birth of CTCF binding sites in the Drosophila genome. PLoS Biol. 10:e1001420.
Park J, Xu K, Park T, Yi SV. 2012. What are the determinants of gene expression levels and breadths in the human genome?. Hum Mol Genet. 21:46–56.
Philippe H, Douady CJ. 2003. Horizontal gene transfer and phylogenetics. Curr Opin Microbiol. 6:498–505.
Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 12:32–42.
Popadin KY, et al. 2014. Gene age predicts the strength of purifying selection acting on gene expression variation in humans. Am J Hum Genet. 95:660–674.
Rainey PB. 2007. Unity from conflict. Nature 446:616–616.
Tautz D, Domazet-Loso T. 2011. The evolutionary origin of orphan genes. Nat Rev Genet. 12:692–702.
Teichmann SA, Babu MM. 2004. Gene regulatory network growth by duplication. Nat Genet. 36:492–496.
Toll-Riera M, et al. 2009. Origin of primate orphan genes: a comparative genomics approach. Mol Biol Evol. 26:603–612.
Uhlen M, et al. 2015. Proteomics. Tissue-based map of the human proteome. Science 347:1260419.
Vishnoi A, Kryazhimskiy S, Bazykin GA, Hannenhalli S, Plotkin JB. 2010. Young proteins experience more variable selection pressures than old proteins. Genome Res. 20:1574–1581.
Wang D, et al. 2012. IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. Bioinformatics 28:729–730.
Webster MT, Hurst LD. 2012. Direct and indirect consequences of meiotic recombination: implications for genome evolution. Trends Genet. 28:101–109.
Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proc Natl Acad Sci U S A. 106:7273–7280.
Xie C, et al. 2012. Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. Plos Genet. 8(9): e1002942.
Yang HW, et al. 2015. Expression profile and gene age jointly shaped the genome-wide distribution of premature termination codons in a *Drosophila melanogaster* population. Mol Biol Evol. 32:216–228.
Zhang JZ. 2003. Evolution by gene duplication: an update. Trends Ecol Evol. 18:292–298.
Zhang WY, Landback P, Gschwend AR, Shen BR, Long MY. 2015. New genes drive the evolution of gene interaction networks in the human and mouse genomes. Genome Biol. 16:202.
Zhang YE, Vibranovski MD, Landback P, Marais GAB, Long MY. 2010. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. PLoS Biol. 8:e1000494.

Zhang Z, et al. 2006. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics 4:259–263.
Zhang Z, et al. 2012. Codon Deviation Coefficient: a novel measure for estimating codon usage bias and its statistical significance. BMC Bioinformatics 13:43.
Zmasek CM, Eddy SR. 2002. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. Bmc Bioinformatics 3:14.

Associate editor: Takashi Gojobori