Historical Constraints on Vertebrate Genome Evolution

Michel C. Milinkovitch*,[1], Raphaël Helaers[2], and Athanasia C. Tzika[1,3]

[1]Laboratory of Artificial and Natural Evolution, Department of Zoology and Animal Biology, Genève, Switzerland

[2]Department of Biology, Facultés Universitaires Notre-Dame de la Paix, rue de Bruxelles, Belgium

[3]Evolutionary Biology & Ecology, Université Libre de Bruxelles, Brussels, Belgium

*Corresponding author: E-mail: michel.milinkovitch@unige.ch.

Accepted: 15 December 2009 **Associate editor:** Wen-Hsiung Li

Abstract

Recent analyses indicated that genes with larger effect of knockout or mutation and with larger probability to revert to single copy after whole genome duplication are expressed earlier in development. Here, we further investigate whether tissue specificity of gene expression is constrained by the age of origin of the corresponding genes. We use 38 metazoan genomes and a comparative genomic application system to integrate inference of gene duplication with expression data from 17,503 human genes into a strictly phylogenetic framework. We show that the number of anatomical systems in which genes are expressed decreases steadily with decreased age of the genes' first appearance in the phylogeny: the oldest genes are expressed, on average, in twice as many anatomical systems than the genes gained recently in evolution. These results are robust to different sources of expression data, to different levels of the anatomical system hierarchy, and to the use of gene families rather than duplication events. Finally, we show that the rate of increase in gene tissue specificity correlates with the relative rate of increase in the maximum number of cell types in the corresponding taxa. Although subfunctionalization and increase in cell type number throughout evolution could constitute, respectively, the proximal and ultimate causes of this correlation, the two phenomena are intermingled. Our analyses identify a striking historical constraint in gene expression: the number of cell types in existence at the time of a gene appearance (through duplication or de novo origination) tends to determine its level of tissue specificity for tens or hundreds of millions of years.

Key words: gene gain, duplication, expression, genome content, phylogeny, metazoa.

Introduction

Recently, Roux and Robinson-Rechavi (2008) have used zebrafish microarray and mouse expressed sequence tag (EST) data spanning, respectively, 14 and 26 developmental stages to investigate whether the timing of expression during development constrains genes' "evolvability." They showed that, in both species, genes with larger effect of knockout or mutation and with larger probability to revert to single copy after whole genome duplication are expressed earlier in development. Their analysis suggests that constraints are high in early stages of vertebrate development and decrease in a monotonous manner over developmental time. Here, we investigate whether these developmental and genomic constraints could be associated to the age of origin of the corresponding genes. Such an analysis requires integrating duplication events and expression data into a strictly phylogenetic framework. Even though phylogeny-based orthology/paralogy identification is widely accepted as the most valid approach (Li et al. 2003; Alexeyenko et al. 2006; Gabaldon 2008; Vilella et al. 2009), many of the methods and databases available for identifying duplication events avoid the heavy computational cost of phylogenetic tree inference and the difficulties associated with their interpretation and, hence, can generate dubious orthology relationships of genetic elements among genomes. Fortunately, more recent databases such as ENSEMBL (Hubbard et al. 2007, 2008) and the "Phylome" approach (Huerta-Cepas et al. 2007, 2008) constitute automated pipelines in which orthologs and paralogs are identified through the estimation of gene family phylogenetic trees. Furthermore, the recently developed MANTiS relational database (Tzika et al. 2008) integrates phylogeny-based orthology/paralogy assignments with functional and expression data, allowing users to explore phylogeny-driven (focusing on any set of branches), gene-driven (focusing on any set of genes), function/process-driven, and expression-driven questions in an explicit phylogenetic framework. We used MANTiS for

assessing the expression constraints of duplicates in relation with their age of origin.

Materials and Methods

Data Mining Data mining and the construction of the relational database were performed with the MANTiS (v1.0.15) pipeline (Tzika et al. 2008), available at www.mantisdb.org and at www.lanevol.org. MANTiS performs automated downloads from ENSEMBL (www.ensembl.org), extracts information relevant to the protein family trees from the Compara database (Vilella et al. 2009), and defines characters for the generation of a full dataset that includes orthologous gene presence/absence information for all selected species. Note that orthology is not assigned on the basis of simple best reciprocal Blast hits (which do not guarantee that orthology is correctly inferred [Theissen 2002] because it ignores gene loss and differential rates of evolution) but through the use of a pipeline that includes 1) the identification of gene families, 2) tree inference after multiple protein sequence alignment within each gene family, and 3) identification of duplication and speciation events through gene tree versus species tree reconciliation. See Tzika et al. (2008) for additional details.

Character Mapping Gains and losses of orthologs are mapped by MANTiS v1.0.15 (www.mantisdb.org and www.lanevol.org) on the "true" species tree (i.e., the topology best supported in [Halanych 2004; Springer et al. 2004; Bashir et al. 2005]). MANTiS maps characters as follows: 1) the character presence/absence matrix for all species (built in the character-mining phase; see above) is used for computing a distance matrix following a modified Jukes–Cantor model, 2) the distance matrix is used to compute branch lengths on the true species topology, using the least-squares approach under minimum evolution, 3) the gain of a character is assigned to the corresponding internal or tip branch of the true species tree, and 4) a recursive maximum likelihood approach is used to identify, for each character, the exact most likely combination of branches on which gene losses are assigned. Once gains and losses have been mapped, MANTiS builds the genome content of each internal node. See Tzika et al. (2008) for much additional details on the character mapping method and genome content view of MANTiS.

Gene Expression Three sources of gene expression data are used in MANTiS: 1) the eGenetics database, using ESTs annotated with eVOC ontology terms (Kelso et al. 2003); 2) the Genomics Institute of the Novartis Research Foundation (GNF) database, including Affymetrix HG-U95A microarray data from the normal physiological state of 25 independent and nonredundant human and 45 mouse tissue samples (Su et al. 2002); and 3) the Human and Mouse Differentially Expressed Genes (HMDEG) database that classifies more than 8 million human and mouse ESTs into tissue/organ categories (Pao et al. 2006). The eVOC anatomical systems include the following 12 first-level categories: "nervous," "urogenital," "alimentary," "respiratory," "endocrine," "cardiovascular," "dermal," "embryo," "musculoskeletal," "hematological," "lymphoreticular," and "unclassifiable." Each of them includes 1–5 lower-level subcategories (e.g., the first-level category "nervous" includes, among others, the following series of hierarchical levels: "central nervous system" → "brain" → "cerebrum" → "cerebral cortex" → "frontal lobe."

Results

Data Mining On the basis of the 38 metazoan genomes (longest splice variant of each protein-coding gene) available in version 49 of the ENSEMBL database (i.e., 6 primates, 1 tree shrew, 4 rodents, 2 lagomorphs, 2 carnivores, 1 perissodactyl, 1 cetartiodactyl, 1 bat, 2 insectivores, 1 xenarthran, 2 afrotherians, 1 marsupial, 1 monotreme, 1 bird, 1 amphibian, 5 teleost fishes, 2 urochordates, 1 nematode, and 3 insects) and the baker's yeast as an outgroup, we used MANTiS v1.0.15 (www.mantisdb.org and www.lanevol.org) to generate two datasets including information on the presence/absence of genes. The first dataset ("families only") contains one character for each single (species specific) gene and for each protein family (i.e., only de novo gains are considered), whereas in the second dataset ("with duplications"), a new character is created for each duplication event, such that each protein family is represented by several characters. More details are given in (Tzika et al. 2008).

Character Mapping Using MANTiS, we mapped gains and losses of characters on the true species phylogeny (i.e., the topology best supported by previous phylogenetic analyses; Halanych 2004; Springer et al. 2004; Bashir et al. 2005): gains are assigned directly from the topology of gene family trees, whereas the most likely positions of gene losses are estimated using a maximum likelihood function (Tzika et al. 2008). These character mapping analyses show that acquisition of new genes through duplication is an important, continuous, and general phenomenon and explains part of the increase of genome size in evolution.

Expression Patterns Three sources of human gene expression data are used in MANTiS: 1) the eGenetics ESTs database (Kelso et al. 2003), 2) the GNF microarray database (Su et al. 2002), and 3) the HMDEG ESTs database (Pao et al. 2006). For each database, MANTiS integrates expression information into categories, representing eVOC ontology terms (Kelso et al. 2003). Expression data are available in the eGenetics database for 16,943 (52.03%) of the human genes. To investigate the level of tissue

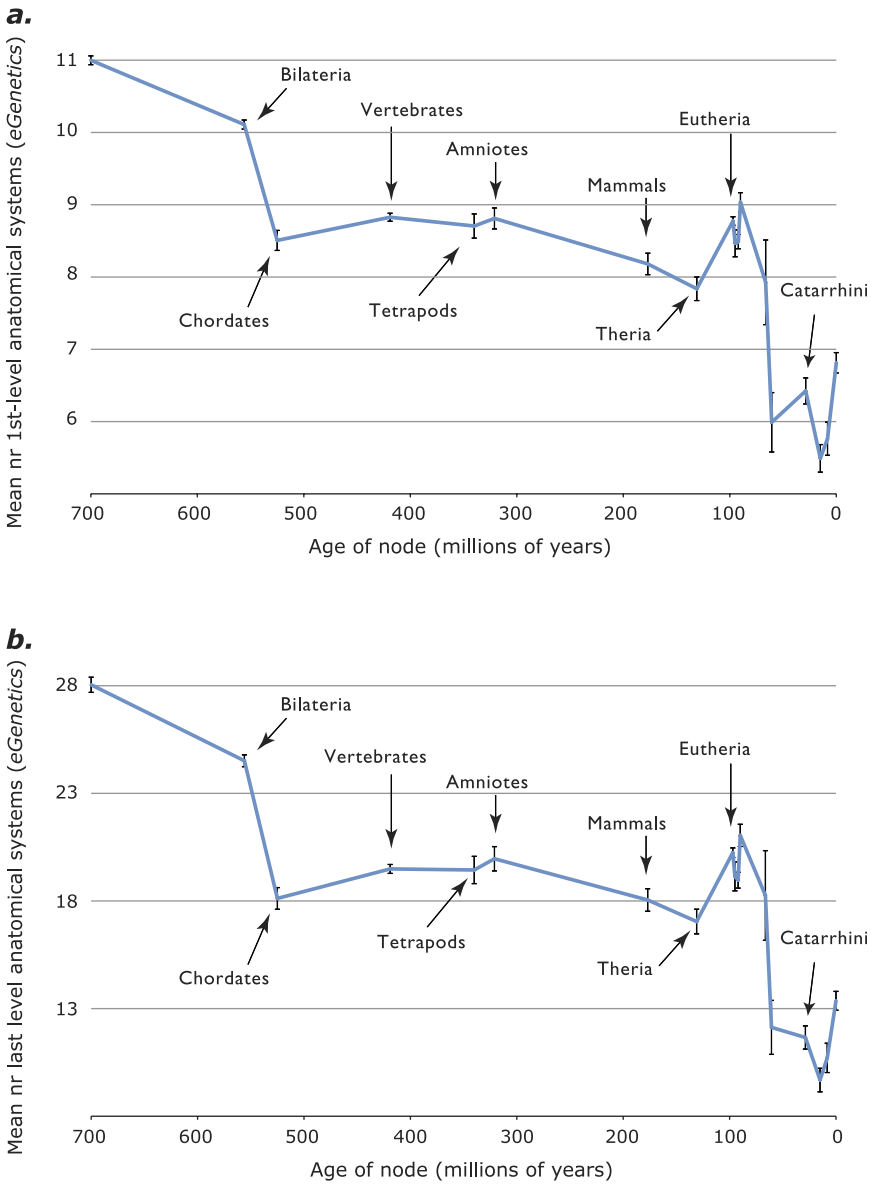

Fig. 1.—Mean number (±standard error) of first-level (*a*) and last-level (*b*) anatomical systems in which human genes are expressed (16,943 genes with available eGenetics expression data) as a function of their first appearance in the phylogeny.

specificity of genes gained along the human lineage, we plotted for all these 16,943 human genes the mean number of first-level anatomical systems in which they are expressed as a function of their first appearance in the phylogeny (fig. 1*a*). This analysis indicates that older genes are less tissue specific: the oldest genes are expressed, on average, in about twice as many anatomical systems than the genes gained recently in evolution. Note that using the last (i.e., deepest)-level anatomical systems does not significantly change the observed pattern (fig. 1*b*). The result is also robust to variations in the expression data source: a very similar decrease in tissue specificity of genes as a function of their age is observed when using the HMDEG ESTs database, both when considering first- or last-level anatomical systems (fig. 2*a* and *b*). Finally, the pattern of change in tissue specificity is even more regular when considering the origin of whole gene families rather than the origin of duplicates (fig. 3, red curve).

This striking pattern might have been brought about by various, nonmutually exclusive mechanisms including 1) broadening of gene expression through evolutionary time, 2) a tendency for duplicates to subfunctionalize, and 3) the differentiation of an increasing number of cell types and anatomical systems through evolutionary time. The latter hypothesis is the simplest. Indeed, the maximum number of somatic cell types (but combining all nerve cell types into a single-cell category) observed in metazoa ranges from four in placozoan to more than 200 in Hominidae and seems to

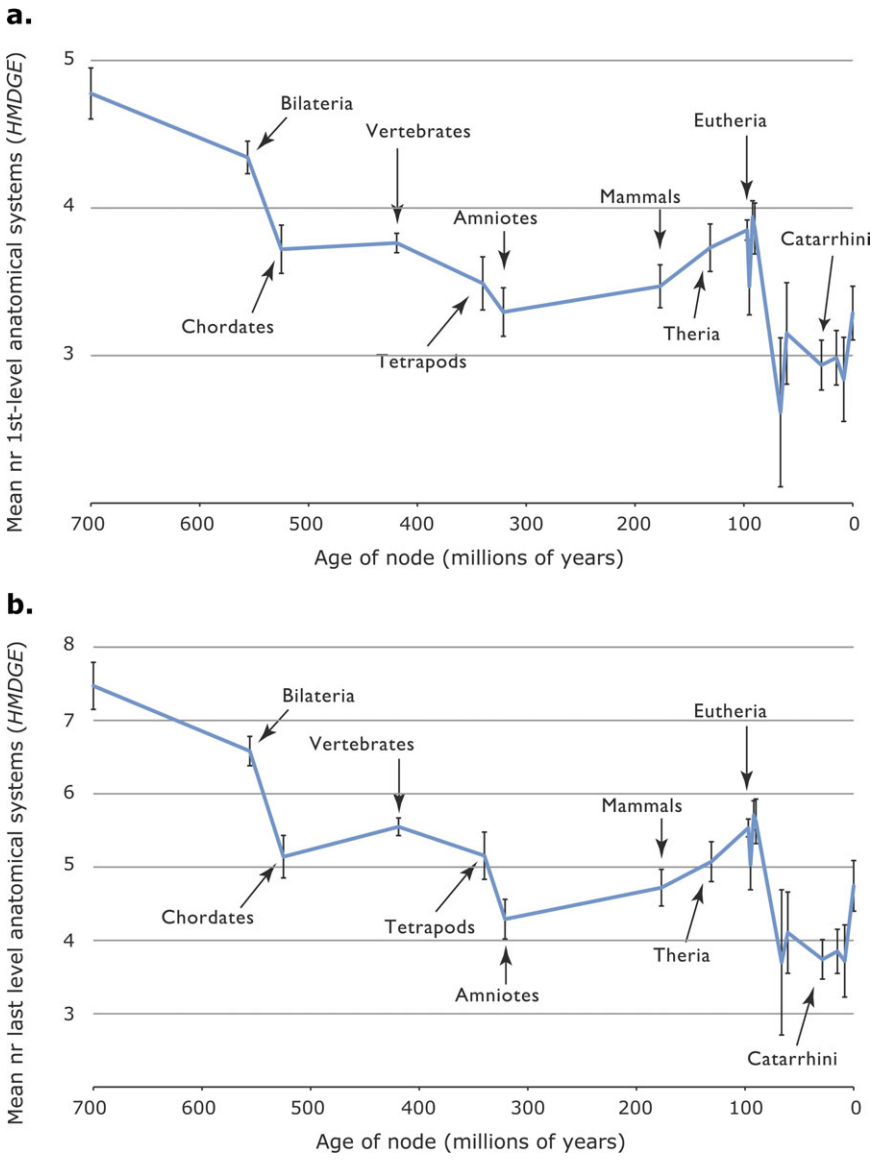

Fig. 2.—Mean number (±standard error) of first-level (*a*) and last-level (*b*) anatomical systems in which human genes are expressed (6,585 genes with available HMDEG expression data) as a function of their first appearance in the phylogeny.

have steadily increased at an average rate of about 0.33 cell type per million years (Valentine et al. 1994). In figure 3, we show that the rate of increase in gene tissue specificity correlates with the relative rate of increase in the maximum number of cell types in the corresponding taxa.

Subfunctionalization is a process by which duplicates can specialize to perform complementary/compartmented functions, hence increase their tissue specificity through protein sequence changes and/or evolution of their respective regulatory modules (Force et al. 1999; Greer et al. 2000; Lynch and Conery 2000; Lynch and Force 2000; Hoekstra and Coyne 2007). Subfunctionalization and the increase in maximum number of cell/tissue types are intermingled: subfunctionalization both 1) requires different cell types and 2) can generate new cell phenotypes, hence an increased number of cell types. Obviously, the development of new cell/tissue types cannot explain alone the pattern of increase in tissue specificity of genes as a function of their age of origin (figs. 1, 2, and 3): subfunctionalization is required. Anyhow, whatever is the timing and relative importance of causal mechanisms, our analysis strongly suggests that the age of first appearance of a gene in the phylogeny is highly predictive of its current level of tissue specificity.

Note that outliers in this general trend are associated with a specific subset of anatomical systems: among the 3,231 genes (with expression data) that originated in the three first branches of the animal phylogeny (the fugi/metazoa, bilateria, and chordates nodes), only 54 are tissue specific (i.e.,

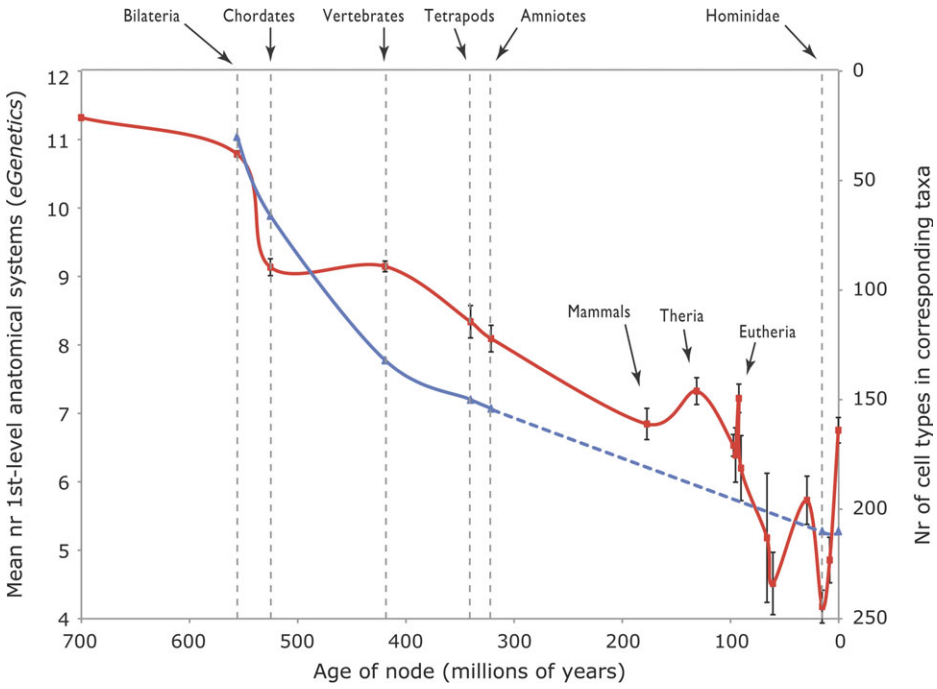

Fig. 3.—The number of cell types in existence at the time of appearance of a gene seems to constrain its level of tissue specificity for hundreds of millions of years. Red line (and primary vertical axis): mean number of first-level anatomical systems in which members of human gene families are expressed (16,943 genes, corresponding to 10,302 families, with available eGenetics expression data) as a function of the family's first appearance in the phylogeny. Blue line (and secondary vertical axis): estimated maximum number of cell types of primitive members of metazoa taxa (indicated with vertical dotted lines). The dashed blue line indicates the gap in available estimates of cell type numbers between early Amniotes and Hominidae. Note that values on the secondary vertical axis are in reverse order.

expressed in a single first-level anatomical system), the majority of which are associated with the urogenital (31 genes), nervous (11 genes), and alimentary (5 genes) anatomical systems (the 7 remaining genes are distributed in 6 of the 8 remaining categories).

Discussion and Conclusions

Morphological novelties abound in the history of animal evolution but increase of complexity and acquisition of novelties are not homogeneously distributed across the phylogenetic tree of life. Although morphological evolution might have been partly driven by the evolution of *cis*-regulatory modules (Carroll et al. 2005), there is little doubt that gene duplications and adaptive structural mutations in protein-coding genes have both contributed substantially to evolution of forms and physiologies (see, e.g., references in Li 1997; Hoekstra and Coyne 2007). Hence, we think that one of the biggest challenges of comparative genomics lies in the identification of changes in genome content that had significant functional implications. Such endeavor may become possible by the integration of genome content and functional data into an explicit phylogenetic framework (Tzika et al. 2008) and should complement 1) analyses of evolutionary conservation (e.g., the characterization of ultraconserved nongenic sequences; Dermitzakis et al. 2003; Bejerano et al. 2004) and 2) identification of protein-coding genes experiencing accelerated sequence evolution (e.g., Clark et al. 2003).

The systematic phylogenetic mapping of gene gains and losses and associated functional data should also prove complementary to the screening of gene expression in target structures at specific stages of their development. Indeed, the latter approach requires prior identification of structures and genes of interest such that it has so far remained mostly restricted to morphological (vs. physiological, metabolic, etc.) characters and to genes known to be likely involved in the development of these structures. Furthermore, these methods of observing spatiotemporal patterns of gene expression do not prove a causal relationship between gene expression and phenotype (Hoekstra and Coyne 2007). The comparative genomic approach on the other hand will require highly accurate genome sequence information and their exhaustive annotation.

Given that homology among genes is inevitably assessed through sequence similarity, different gene families might actually represent a single gene family that has been artificially split. Indeed, old duplication events can have generated subfamilies whose divergence observed today exceeds the dissimilarity thresholds used in homology inference methods. In other words, some gains inferred as de novo gene gains in MANTiS might correspond to duplication events.

This phenomenon is very unlikely to have any impact on our conclusions because the trend (increased tissue specificity with decreased age of gene origin) is observed when we include both de novo gains and duplication events (i.e., with the dataset "with duplications," figs. 1 and 2), when we include de novo gains only (i.e., with the dataset "families only," fig. 3), and when we include duplication events only (data not shown). Finally, our preliminary analyses of mouse expression data indicate a very similar trend (data not shown) as with human data.

In conclusion, despite low sequence coverage of several "full" genomes, substantial imperfections in genome annotation, and a large taxonomic bias in the species whose genomes have been sequenced, our integrated analyses of expression and genome content data in a strictly phylogenetic framework identify a striking historical constraint in gene expression: the number of cell types in existence at the time of appearance of a gene constrains its level of tissue specificity for tens or hundreds of millions of years. Testing whether this hypothesis is generalizable would require similar analyses along other lineages, for example, of nonchordates, that is, branches which diverged early from the lineage shown in figures 1–3. Ultimately, expression data from multiple lineages should be incorporated, such that ancestral states of tissue specificity would be inferred for each gene at each node of the phylogeny.

Acknowledgments

This work was supported by funds from the University of Geneva (Switzerland), the Swiss National Science Foundation (FNSNF, grant 31003A_125060), the Société Académique de Genève, the Georges and Antoine Claraz Foundation, and the Ernst and Lucie Schmidheiny Foundation. We are grateful to Andreas Wagner and two anonymous reviewers for their constructive comments on previous versions of the manuscript.

Literature Cited

Alexeyenko A, Tamas I, Liu G, Sonnhammer EL. 2006. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics. 22:e9–e15.

Bashir A, Ye C, Price AL, Bafna V. 2005. Orthologous repeats and mammalian phylogenetic inference. Genome Res. 15:998–1006.

Bejerano G, et al. 2004. Ultraconserved elements in the human genome. Science. 304:1321–1325.

Carroll SB, Grenier JK, Weatherbee SD. 2005. From DNA to diversity, molecular genetics and the evolution of animal design. Malden (MA): Blackwell publishing.

Clark AG, et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science. 302:1960–1963.

Dermitzakis ET, et al. 2003. Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). Science. 302:1033–1035.

Force A, et al. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 151:1531–1545.

Gabaldon T. 2008. Large-scale assignment of orthology: back to phylogenetics? Genome Biol. 9:235.

Greer JM, Puetz J, Thomas KR, Capecchi MR. 2000. Maintenance of functional equivalence during paralogous Hox gene evolution. Nature. 403:661–665.

Halanych KM. 2004. The new view of animal phylogeny. Annu Rev Ecol Evol Syst. 35:229–256.

Hoekstra HE, Coyne JA. 2007. The locus of evolution: evo devo and the genetics of adaptation. Evol Int J Org Evol. 61:995–1016.

Hubbard TJ, et al. 2008. Ensembl 2009. Nucleic Acids Res. 37:D690–D697.

Hubbard TJ, et al. 2007. Ensembl 2007. Nucleic Acids Res. 35: D610–D617.

Huerta-Cepas J, Bueno A, Dopazo J, Gabaldon T. 2008. PhylomeDB: a database for genome-wide collections of gene phylogenies. Nucleic Acids Res. 36:D491–D496.

Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldon T. 2007. The human phylome. Genome Biol. 8:R109.

Kelso J, et al. 2003. eVOC: a controlled vocabulary for unifying gene expression data. Genome Res. doi: 10.1101/gr.985203 13:1222–1230.

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13: 2178–2189.

Li W-H. 1997. Molecular evolution. Sunderland (MA): Sinauer.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science. 290:1151–1155.

Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics. 154:459–473.

Pao S-Y, Lin W-L, Hwang M-J. 2006. In silico identification and comparative analysis of differentially expressed genes in human and mouse tissues. BMC Genomics. 7:1–11.

Roux J, Robinson-Rechavi M. 2008. Developmental constraints on vertebrate genome evolution. PLoS Genet. 4:e1000311.

Springer MS, Stanhope MJ, Madsen O, de Jong WW. 2004. Molecules consolidate the placental mammal tree. Trends Ecol Evol. 19: 430–438.

Su AI, et al. 2002. Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci USA. 99:4465–4470.

Theissen G. 2002. Secret life of genes. Nature. 415:741.

Tzika A, Helaers R, Van de Peer Y, Milinkovitch MC. 2008. MANTiS: a phylogenetic framework for multi-species genome comparisons. Bioinformatics. 24:151–157.

Valentine JW, Collins AG, Meyer CP. 1994. Morphological complexity increase in metazoans. Paleobiology. 20:131–142.

Vilella AJ, et al. 2009. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 19:327–335.