How confident can we be that orthologs are similar, but paralogs differ?

Romain A. Studer[1,2] and Marc Robinson-Rechavi [1,2,3]

[1] Department of Ecology and Evolution, Biophore, Lausanne University, CH-1015 Lausanne, Switzerland;

[2] Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland;

[3] Corresponding author: marc.robinson-rechavi@unil.ch; fax 41-21-6924165.

Abstract

Homologous genes are classified into orthologs and paralogs, depending on whether they arose by speciation or duplication. It is widely assumed that orthologs share similar functions, whereas paralogs are expected to diverge more from each other. But does this assumption hold up on further examination? We present evidence that orthologs and paralogs are not so different in their evolutionary rates nor in their mechanisms of divergence. We emphasize the importance of appropriately designed studies to test models of gene evolution between orthologs and between paralogs. Thus functional change between orthologs may be as

common as between paralogs, and future studies should be designed to test the impact of duplication against this alternative model.

The relationship between gene duplication and gene function

Understanding how genes acquire new functions is necessary if we are to have a more complete understanding of molecular evolution. Particular attention has been given to gene duplication because it is often assumed that changes in gene function are preferentially associated with duplication [1]. This means that it is important to distinguish orthologs from paralogs (Figure 1) [2, 3], because experimental information concerning one gene should be readily generalized to all its orthologs. The distinction between orthology and paralogy has been emphasized in recent years through the revival of interest in gene duplication [1, 4], and the emergence of comparative approaches as a major tool of genomics [5]. The idea that orthologs share similar functions, while paralogs have different functions, has thus become accepted by many and is the standard textbook model, as exemplified by the "Phylogenetics Factsheet" of NCBI (http://www.ncbi.nlm.nih.gov/About/primer/phylo.html). Of note, only one-to-one orthologs should be expected to conserve function if duplication has an important impact, a distinction made in some resources [6, 7].

The focus on duplication has also led to the elaboration of theoretical models of evolution after duplication and their testing in genome wide studies [1, 8]. It is recognized that duplication does not always lead to changes in function. But the assumption that changes in function are generally associated with duplication has rarely been explicitly tested. While there have been many studies of comparative genomics focused on the role of duplication (reviewed in [1]), few have compared the evolution of paralogs with the evolution of

orthologs. However, these studies repeatedly find little if any specific impact of duplication. This pattern is surprising if the standard model is correct.

This "standard model" makes two predictions. First, paralogs are expected to diverge more, per unit of time, than orthologs. Second, paralogs are expected to diverge frequently in ways that are rarely observed between orthologs; for example, changes in substrate specificity. Divergence can concern different aspects of gene function [3], such as constraints on protein sequence or structure, patterns of expression, or participation in molecular networks. We contrast this to an "alternative model", under which all homologs diverge approximately proportionally to time, whether they are paralogs or orthologs. First, distant homologs are expected to differ more than close homologs; notably, recent paralogs should have more similar function than ancient orthologs. Second, radical changes, e.g. in substrate specificity, are expected to be rare both between paralogs and between orthologs, but increasingly probable with time of divergence. This view is partially captured by the concept of in-paralogs [3, 9] (Figure 1), according to which paralogs which have diverged recently are expected to share similar functions. But no change is still expected between orthologs. And continuous divergence of paralogs over time is not fully represented by two discrete classes, defined by a unique speciation event.

We examine the evidence that the evolution of gene function is different during the divergence of paralogs and orthologs. Our aim here is to highlight open questions, not to provide a comprehensive review on the nature of gene duplication. For recent reviews, please see [1, 8, 10]. We emphasize that the design of comparative genomics studies has a major impact on the conclusions which can be drawn. Comparing paralogs within a species can provide a measure of divergence after duplication, but cannot prove a specific role for duplication. Thus although many studies report divergence between paralogs, their design does not usually enable the standard model to be tested.

3

Asymmetric rates of sequence evolution within and between species

One of the main themes of genomic studies of duplicate genes is that the copies do not evolve symmetrically [8]; that is they do not evolve at the same rate. Protein sequences are known to evolve at different rates between paralogs [11-13], providing a rough indication of the evolution of biochemical function. The rate of evolution of protein sequences has also been abundantly studied between orthologs, where asymmetry is frequent. For example, proteins evolve faster in rodents than in other mammals [14], and rates are variable among insects [15]. More generally, the hypothesis of a constant rate of evolution rarely seem to hold [16], which provides indirect evidence for a widespread asymmetry of rates of evolution. To our knowledge, the level of asymmetry has not been directly compared between paralogs and orthologs in a unified study.

The explanations for differences in the rates of evolution between species (i.e. when examining orthologs) have focused on life history traits, such as generation time and population size [14], whereas comparisons of paralogs have focused on functional change [1]. While paralogs in the same species are not expected to differ in life history traits, asymmetry can be affected by non functional differences between paralogs, such as differences in recombination [17]. And importantly, functional change could affect orthologs.

To take a "gene's eye" view, the important distinction is between two mechanisms of increase in the rate of evolution: (i) a relaxation of purifying selection, and (ii) an increase in positive selection. Both mechanisms can cause changes in gene function. And both can result from changes in population size, functional constraints, or other causes. Thus from this perspective, the evolution of paralogs and of orthologs can be affected by the same mechanisms.

Divergence in gene expression

The expression levels of paralogs also evolve asymmetrically, as reported in yeast [18, 19], Arabidopsis [11, 20] and Xenopus [21]. Comparing gene expression levels between different species is difficult, because of differences in established experimental conditions, and in their anatomy or life cycles. For this reason, the putative regulatory elements are also often examined: if the regulatory sequence changes, then it is likely that the expression of the gene is also altered [22]. Thus asymmetric divergence of *cis*-elements has been reported in teleost fish [23] and in yeast [19].

The extent of divergence in expression after duplication has rarely been compared with divergence in expression after speciation on a large scale. In general, duplicate genes show faster divergence of expression than singleton genes, as reported in mammals [24] and Drosophila [25]. Faster divergence for duplicate genes is also found for the divergence of *cis*-regulatory sequences in nematodes [22]. But the design of these studies cannot distinguish whether (i) genes whose expression evolves rapidly are retained in duplicate, or (ii) gene expression evolves faster after duplication. This also highlights the importance of distinguishing one-to-one orthologs from orthologs with secondary duplication, i.e. in-paralogs (Figure 1).

It has been suggested that functional divergence after duplication is mostly due to changes in expression [19]. Curiously, the same claim has been made for divergence between species [26], and genomic studies have reported significant divergence of expression patterns between orthologs. One of the most conservative studies, a comparison of human and mouse correcting for experimental differences between species and for estimation error [27], reports 16% of orthologs whose expression appears to diverge neutrally, while another third diverge in a detectable manner (d > 0.02 in Fig. 4 of ref [27]) despite purifying selection.

Changes in protein function

Although global changes in rates of protein evolution inform us that the process is not constant, they are not very informative about specific changes in function [3, 28, 29]. For that, either specific proteins must be investigated in detail, or more specific classes of amino acid changes need to be compared. Small scale studies have shown that functional divergence occurs both between orthologs and between paralogs, although they cannot provide a test for the relative importance of these events.

The nuclear hormone receptor superfamily provides classical examples of functional divergence between paralogs, but divergence between orthologs can also be found. For example, steroid receptors appear to have developed specialized functions after duplication [30]. Yet in amphioxus the ortholog of vertebrate estrogen receptor does not bind to estrogen, whereas the ortholog of other steroid receptors does [31]; in this case paralogs share function, not orthologs. Nuclear receptor function can also change in the absence of any duplications. For example the Drosophila ecdysone receptor differs from other insect orthologs in ligand binding and in dimerization pattern [32]. More generally, there is mounting evidence that orthologous transcription factors are not always functionally equivalent [33].

Other well studied protein families show a similar pattern, with some functional changes between paralogs, and others between orthologs. This is for example the case of the remarkable shifts in wavelength sensitivity between *M/LWS* (Middle/Long Wavelength Sensitive) pigments [34]. Changes in enzymatic activity have also been reported between orthologs, for example between lysozymes, with adaption to herbivorous diets [35]; or between RubisCO enzymes in plants, with convergent adaptation to dry conditions [36].

Large-scale amino acid studies

Most small scale studies show that changes in a few key protein sites have resulted in a change in function. More generally, such changes in biochemical function can result from

either (i) a rare change in amino acid at a site that remains constrained under its new form; or (ii) a change in selective pressure ("covarions"); that is sites with different evolutionary rates in different parts of the phylogeny, due to changes in functional role [28, 37, 38]. For example, residue W348 is invariant in *CED-3* caspases, where it is critical for substrate specificity, whereas it is highly variable in paralogous *ICE* caspases [39]. Several methods to detect such changes have been developed explicitly to study duplication [37, 40]. But studies of orthologs have also found evidence for "covarions", for example in functionally important sites between HIV-1 subtypes [41].

In the vertebrate hemoglobin family, similar proportions of "covarion" sites were reported between orthologs and paralogs, but an excess of changes in constrained amino acids was found between paralogs [42]. This excess of rare changes was confirmed in a larger study of proteins domains [43], and provides some evidence for specific divergence after duplication. No large scale study of "covarions" has been done to our knowledge.

Amino-acid changes can also be classified as radical (e.g. a change in physicochemical properties) or conservative. Applying a simple model of evolution to 1821 proteins domains, the excess of radical replacement after duplication was found to be non significant relative to speciation [43]. Similarly, no difference in the type of amino acid changes was found after whole genome duplication in yeast [13]. The application of more elaborate models to estimate radical changes in amino acids in five families of orthologous genes and five of paralogous genes found high variability in the substitution process, but again no difference between paralogs and orthologs was found [44]. This similarity of patterns of amino-acid substitutions between paralogs and orthologs implies that evolution of the molecular function of proteins may not follow the standard model of divergence after duplication.

Positive selection

A reasonable simplifying assumption to detect neofunctionalization is that it is driven at least in part by positive selection [1]. Thus the detection of branch specific episodes of positive selection could be indicative of changes in protein function after duplication [45, 46]. In snake venom *PLA₂* genes, positive selection has been linked with the evolution of toxin function both after duplication and after speciation [47]. We have reported recently a scan for such positive selection in vertebrate genes [48]. While we found that evolution of one third of paralogs has been shaped by positive selection after duplication, this was not more than detected in the absence of duplication. We did find an excess of relaxed purifying selection after duplication. This relaxation may explain the patterns of divergence of constrained amino acid sites between paralogs [42, 43], and of high asymmetry between paralogs [13].

Interestingly, positive selection has been increasingly detected between orthologs with the recent progress of data and methods [49-51]. The functional implications of these observations are not yet clear, but such selection appears weak, although frequent [48, 51, 52]. It could thus correspond to an accumulation of small changes, which may or may not result in large functional change over time [50]. An integration of the role of frequent positive selection between orthologs into the standard model (i.e. that most functional innovations are associated with duplication) remains to be performed.

The importance of study design in evolutionary genomics

An important issue for evaluating the impact of duplication on gene function is study design. Obviously, if only divergence between paralogs is inquired into, then divergence between orthologs cannot be reported (Figure 2, i–iii; Table 1). The major role of design in the study of paralogs has previously been highlighted by the demonstration that a biased subset of genes is retained in duplicate [53]. This "Davis and Petrov effect" [13] can lead to confusion

between the effect of duplication on genes and the effect of gene retention. Thus, we need to control for biased retention to conclude there is an effect of duplication, (e.g. [12, 13]; Figure 2, iv–v). Furthermore, we need to control for evolution in the absence of duplication (e.g. [43, 48]; Figure 2, viii-ix) to conclude that duplication has a *specific* effect. Because of biased gene retention, this control should include singleton orthologs of the duplicated genes.

Which study to test which evolutionary model?

In Table 1 and Figure 2, we propose a classification of the design of studies of duplicate genes, with the predictions they allow under three simplified models of evolution: subfunctionalization after duplication, neofunctionalization after duplication, and the "alternative model" of equal change after duplication or speciation. In this cartoon version of neo-functionalization, one copy evolves strictly in the same manner as a singleton, while the other acquires a new function by positive selection. Similarly, in this version of sub-functionalization, any gain of function is excluded. It is clear that more complex evolutionary fates are possible and probable [1, 54], but our point here is that even such simplistic models, with very different expectations, will not be distinguished by inadequate study design. We represent the "alternative model" by a version of neofunctionalization without the assumption that it occurs preferentially after duplication. It should be noted that this alternative model is voluntarily simplistic, for illustration purposes, and that we do not propose that it effectively describes gene evolution.

It is apparent that some designs do not allow any differential predictions, including the practice of comparing pairs of paralogs to random pairs within the same species (Figure 2, ii). This does not mean of course that such studies are not useful. For example, protein–protein interaction data is available only for a handful of distantly related model organisms, which limits possible study designs. Although this has led to difficulty in defining clear evolutionary trajectories [55], notable results include biased retention of paralogs relative to network

position [19], and that paralogs interact frequently with each other [55]. Such interactions could facilitate the evolution of new functions, for example if one member of a heterodimer pair loses the original function, but retains the ability to interact, becoming a specific repressor [31].

In a more complex design, a small but significant excess of amino acid changes on duplication branches was found in a phylogenomic study of Chordate genes [56]. But curiously, the acceleration concerned both branches preceding and following the duplication. This suggests that in the Chordate dataset analyzed, there may be simultaneous pressure for substitutions and duplication in some "diversifying genes" [56]. That would make it difficult to disentangle retention bias and the effects of duplication, in a design where speciation branches were analyzed without taking into account the rest of the evolutionary history of each gene (i.e. frequent or rare duplications).

Not born equal

But what about the duplication events themselves? Different mechanisms result in the duplication of different functional categories of genes [57]. Moreover, a distinction should be made between mechanisms which are symmetric or not "at birth" [58]: at one extreme, paralogs formed as a result of whole genome duplication are initially redundant in every aspect of their genomic context and organization; whereas at the other extreme, retrotransposed genes differ profoundly from their parent genes as soon as they are generated [10]. This has an obvious impact on the expectations of evolution after duplication.

A more subtle bias is that singleton genes in a lineage that experienced whole genome duplication have undergone a period of evolution as one member of a redundant pair, which might distinguish their evolutionary trajectory from that of orthologs, which did not experience such an event (e.g. mammalian versus. teleost orthologs [59]). Finally, young and old paralogs can differ, because evolutionary pressure may change over time [1], and because different types of genes can be retained in duplicate for different lengths of time. Thus it is to

be expected that different models explain the evolution of paralogs which have been generated by different mechanisms.

Concluding remarks

Few large scale studies have been conducted that allow an explicit testing of all three models presented in Table 1, let alone more elaborate ones. It is clear that, as often in evolutionary biology, specific cases of all scenarios can be found. The questions are thus: what is the most common mode of evolution? and do minority modes of evolution also play an important role, or are they of negligible impact?

Despite limitations, we are struck by the number of small or large scale studies which report less difference than expected under the "standard model" between the evolution of paralogs and of orthologs [2, 3]. In addition, orthologs are increasingly found to diverge without duplication, in sequence [49-51], in expression [27], and even in Knock-Out phenotypes [60]. But the existence of these studies does not seem to modify the standard view of phylogenomics, as summarized in e.g. the NCBI Factsheet. Perhaps this reflects the lack of interest for negative results: if a study did not find a difference, and we "know" that there is one, then surely they did not look well enough, or in the right place? There might be something to this view, in that it is possible that protein sequence evolution, for which we have the most data, is less impacted by duplication than other features [19]. In addition there appears to be consistent support for a relaxation of purifying selection on sequences after duplication, although its impact on function remains to be established.

Whether changes in gene function occur preferentially after duplication or not is important for our understanding of evolution, since duplication is frequently viewed as the preferred mechanism to generate novelty in genomes [1]. It is also important to evaluate the relevance of transferring annotations between orthologs. In this context, an intriguing recent result is that sequence similarity appears to be a better predictor of common Gene Ontology

terms than orthology [61]. This is expected under the "alternative model" of functional divergence with time, but not under the "standard model" of preferential divergence after duplication.

In conclusion, we would like to emphasize two points: one methodological, that the design of studies of comparative genomics imposes strong limitations on the questions which can be answered; and one biological, that changes in function may be as common between orthologs as between paralogs. Future work should focus on testing the role of duplication versus speciation with appropriate designs and data.

Acknowledgements

Glossary

Bias in gene retention: The retention of copies of genes after duplication is not random, relative to gene function or evolutionary rate. This bias can be a confounding factor in large scale analyses of paralogs: the function of duplicate genes in a genome results both from the bias in retention and from evolution after duplication.
Homologs: Genes descending from a common ancestor.
Inparalogs: Paralogs resulting from a duplication after a speciation event of reference.
Negative selection: Selection which decreases the chance of fixation of a mutation, because it is detrimental; it results in an decrease of the rate of evolution of selected mutations.
Neofunctionalisation: The process in which one paralog gains a new function, which is selectively advantageous.
One-to-one orthologs : Orthologs which are present in a single copy in each genome of interest.
Orthologs: Homologs which have diverged since a speciation event.
Outparalogs: Paralogs resulting from a duplication before a speciation event of reference.
Paralogs: Homologs which have diverged since a duplication event.
Positive selection: Selection which increases the chance of fixation of a mutation, because it is beneficial; it results in an increase of the rate of evolution of selected mutations.
Subfunctionalization: The process in which paralogs partition the ancestral function, so that each performs only part of this function. Subfunctionalization may happen by degenerating reciprocal mutations (DDC model), or by positive selection for specialization (escape from adaptive conflict).

Figure 1: Formation of orthologs and paralogs

The evolutionary tree shows six homologous genes from three species designated A, B and C.

Genes are represented by circles and each color represents a different species; genes with

paralogs are circled by a thicker line (only the gene in the A lineage does not have a paralog).

Boxes at nodes represent duplication events. Duplication 1 produced paralogs α and β in the

ancestor of B and C, whereas Duplication 2 produced paralogs β_1 and β_2 in the C lineage. All

genes from B and C are co-orthologs to the gene from A. Genes α and β are in-paralogs

relative to Speciation 1, but out-paralogs relative to Speciation 2. Genes β_1 and β_2 are in-

paralogs relative to both speciations in the tree. Genes Bα and Cα are one-to-one orthologs.

Figure 2: Some designs for the study of gene duplication

Designs i-iii represent three strategies to study paralogs using information from a single

species: (i) measuring the divergence between paralogs; (ii) contrasting the divergence of

paralogs to the divergence of random gene pairs; (iii) contrasting the characteristics of paralogs to those of single copy genes (genes without a paralog detected in the same species). Designs iv-vii represent four ways of using outgroup species to determine more accurately the divergence of paralogs; note that in all cases these are in-paralogs relatively to speciation with the outgroup: (iv) using singleton orthologs to determine the characteristics of genes retained as paralogs; (v) using pairs of singleton orthologs to determine the evolutionary rate of genes retained as paralogs; (vi) determining asymmetry of paralog divergence; (vii) contrasting divergence from the outgroup of paralogs vs. singletons. Designs viii-ix represent a complete phylogenetic analysis, contrasting evolution of orthologs and paralogs in outgroups and ingroups; note that defining in-paralogs and out-paralogs in this case depends on the speciation used as reference, and that the paralogs are all co-orthologs to the singletons: (viii) comparison of branch specific evolutionary rates; (ix) comparison of functional characteristics of genes. The symbols used as the same as in Figure 1. Genes are represented by circles; each color represents a different species; genes with paralogs are circled by a thicker line. Boxes at nodes represent duplication events. The thick broken arrows indicate which elements are compared in order to study the effect of duplication, whereas thin broken arrows indicate other comparisons included in this design. These examples of design show the importance both of using more species, and of defining the phylogenetic relationships between genes under study.

Table 1: The impact of study design on tests of evolution after duplication

Study design [a]	Data type [b]	Examples	Predictions under simple evolutionary models		
			Preferential change after duplication		Function change after duplication or speciation
			subfunctionalization [c]	neofunctionalization	
i- Divergence between duplicates	Functional	[19, 20, 55]	Differences between paralogs		
ii- Duplicate pairs *vs.* random pairs	Functional	[11, 19, 54]	Paralogs more similar than random pairs, but not identical		
iii- Duplicates *vs.* singletons	Functional	[11, 19, 25]	Measure of retention bias, confused by evolution after duplication		
iv- Orthologs of duplicates *vs.* orthologs of singletons	Functional	[12]	Measure of retention bias		
v- Divergence between orthologs of duplicates *vs.* between orthologs of singletons	Sequences	[12, 53]	Measure of retention bias		
vi- Divergence relative to outgroup	Sequences	[11, 17, 58]	No prediction relative to symmetry, relaxed purifying selection	Asymmetry between paralogs, positive selection [e]	
	Functional	[18, 21]	Two paralogs different, complementary to full outgroup function	One paralog similar to outgroup, one different	
vii- Divergence relative to outgroup of duplicates *vs.* singletons	Sequences	[62]	Higher divergence of duplicates, confused by retention bias		
	Functional	[18, 24, 25]	Two paralogs different,	One paralog similar to	No specific prediction [f]

			complementary to outgroup; singleton similar to outgroup	outgroup, one different; singleton similar to outgroup	
viii- Comparison of several singletons and duplicates per gene tree	Sequences	[13, 43, 48, 56]	Higher relaxation of purifying selection on branches after duplication	More positive selection on branches after duplication	Positive selection in various branches of the tree [g]
ix- *idem*	Functional		Conservation of pattern among singletons; sub-patterns in duplicates	Conservation in most homologs; new patterns [h] in some duplicates	Variation in pattern among homologs, with gain of new patterns [h]

(a) The numbering of study designs follows that outlined in Figure 2.

(b) "Functional"; comparison of functional genomics data (e.g. microarrays, protein-protein interactions), or of functional annotations (e.g. Gene Ontology); "Sequence": sequence based comparisons (e.g. dN/dS).

(c) Limited here to DDC subfunctionalization; see Ref [1].

(d) Corrected for divergence time (i.e. per Million years, or per dS).

(e) In principle, positive selection may be expected, but this design provides very little power to detect it.

(f) The outgroup also might have changed under this model.

(g) Without an excess on branches following duplication.

(h) E.g. new expression domain, new interaction partner, not found in other homologs. The strength of inference depends on the number of homologs with functional data.
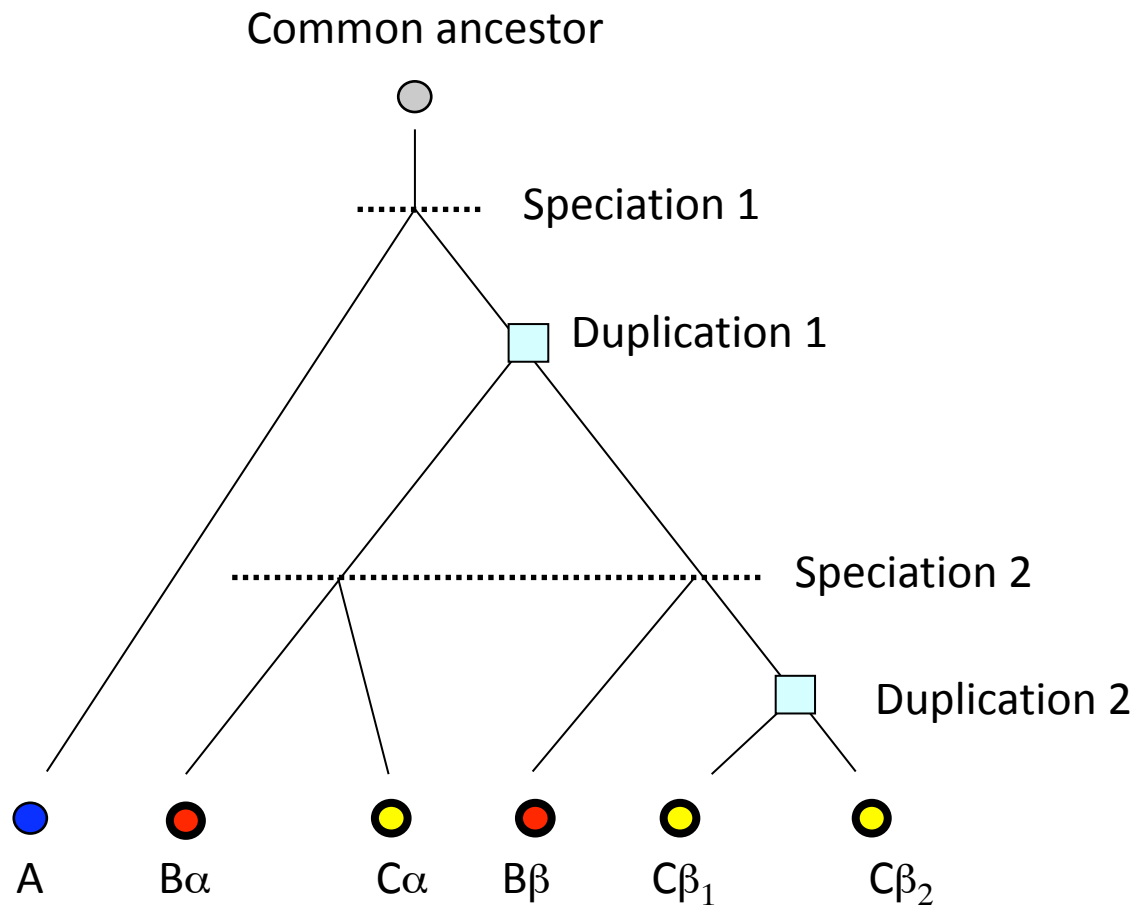
All predictions must be understood as statistical (i.e. applicable to large datasets only).
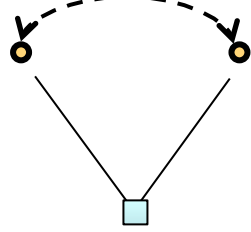
References

1.   Conant, G.C., and Wolfe, K.H. (2008) Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet* 9, 938-950

2.   Koonin, E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 39, 309-338

3.   Rentzsch, R., and Orengo, C.A. Protein function prediction - the power of multiplicity. *Trends Biotechnol* In Press, Corrected Proof

4.   Taylor, J.S., and Raes, J. (2004) Duplication and divergence: The Evolution of New Genes and Old Ideas. *Annu Rev Genet* 38, 615-643

5.   Margulies, E.H., and Birney, E. (2008) Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. *Nat Rev Genet* 9, 303-313

6.   Berglund, A.C.*, et al.* (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res* 36, D263-266

7.   Hubbard, T.J.*, et al.* (2009) Ensembl 2009. *Nucleic Acids Res* 37, D690-697

8.   Semon, M., and Wolfe, K.H. (2007) Consequences of genome duplication. *Curr Opin Genet Dev* 17, 505-512

9.   Sonnhammer, E.L.L., and Koonin, E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet* 18, 619-620

10.   Kaessmann, H.*, et al.* (2009) RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* 10, 19-31

11.   Blanc, G., and Wolfe, K.H. (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* 16, 1679-1691

12.   Brunet, F.G.*, et al.* (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol* 23, 1808-1816

13.   Scannell, D.R., and Wolfe, K.H. (2008) A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Res* 18, 137-147

14.   Nikolaev, S.I.*, et al.* (2007) Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proc Natl Acad Sci U S A* 104, 20443-20448

15.   Zdobnov, E.M., and Bork, P. (2007) Quantification of insect genome divergence. *Trends Genet* 23, 16-20

16.   Graur, D., and Martin, W. (2004) Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends Genet* 20, 80-86

17.   Clement, Y.*, et al.* (2006) Does lack of recombination enhance asymmetric evolution among duplicate genes? Insights from the Drosophila melanogaster genome. *Gene* 385, 89-95

18.   Tirosh, I., and Barkai, N. (2007) Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biol* 8, R50

19.   Wapinski, I.*, et al.* (2007) Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449, 54-61

20.   Ganko, E.W.*, et al.* (2007) Divergence in expression between duplicated genes in Arabidopsis. *Mol Biol Evol* 24, 2298-2309

21.   Semon, M., and Wolfe, K.H. (2008) Preferential subfunctionalization of slow-evolving genes after allopolyploidization in Xenopus laevis. *Proc Natl Acad Sci U S A* 105, 8333-8338

22.   Castillo-Davis, C.I.*, et al.* (2004) cis-Regulatory and protein evolution in orthologous and duplicate genes. *Genome Res* 14, 1530-1536

23.   Woolfe, A., and Elgar, G. (2007) Comparative genomics using Fugu reveals insights into regulatory subfunctionalization. *Genome Biol* 8, R53

24.   Huminiecki, L., and Wolfe, K.H. (2004) Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res* 14, 1870-1879

25.   Gu, Z*., et al.* (2004) Duplicate genes increase gene expression diversity within and between species. *Nat Genet* 36, 577-579

26.   Prud'homme, B*., et al.* (2007) Emerging principles of regulatory evolution. *Proc Natl Acad Sci U S A* 104 Suppl 1, 8605-8612

27.   Liao, B.Y., and Zhang, J. (2006) Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol Biol Evol* 23, 530-540

28.   Philippe, H*., et al.* (2003) Heterotachy and functional shift in protein evolution. *IUBMB Life* 55, 257-265

29.   Kristensen, D*., et al.* (2008) Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. *BMC Bioinformatics* 9, 17

30.   Bridgham, J.T*., et al.* (2006) Evolution of hormone-receptor complexity by molecular exploitation. *Science* 312, 97-101

31.   Bridgham, J.T*., et al.* (2008) Evolution of a new function by degenerative mutation in cephalochordate steroid receptors. *PLoS Genet* 4, e1000191

32.   Iwema, T*., et al.* (2009) Structural and Evolutionary Innovation of the Heterodimerisation Interface between USP and the Ecdysone Receptor ECR in Insects. *Mol Biol Evol* in press

33.   Lynch, V.J., and Wagner, G.P. (2008) Resurrecting the role of transcription factor change in developmental evolution. *Evolution* 62, 2131-2154

34.   Yokoyama, S. (2008) Evolution of Dim-Light and Color Vision Pigments. *Ann Rev Genomics Human Genet* 9, 259-282

35.   Messier, W., and Stewart, C.-B. (1997) Episodic adaptive evolution of primate lysozymes. *Nature* 385, 151-154

36.   Christin, P.A*., et al.* (2008) Evolutionary switch and genetic convergence on rbcL following the evolution of C4 photosynthesis. *Mol Biol Evol* 25, 2361-2368

37.   Gu, X. (1999) Statistical methods for testing functional divergence after gene duplication. *Mol Biol Evol* 16, 1664-1674

38.   Anisimova, M., and Liberles, D.A. (2007) The quest for natural selection in the age of comparative genomics. *Heredity* 99, 567-579

39.   Wang, Y., and Gu, X. (2001) Functional Divergence in the Caspase Gene Family and Altered Functional Constraints: Statistical Analysis and Prediction. *Genetics* 158, 1311-1320

40.   Edwards, R.J., and Shields, D.C. (2005) BADASP: predicting functional specificity in protein families using ancestral sequences. *Bioinformatics* 21, 4190-4191

41.   Penn, O*., et al.* (2008) Evolutionary modeling of rate shifts reveals specificity determinants in HIV-1 subtypes. *PLoS Comput Biol* 4, e1000214

42.   Gribaldo, S*., et al.* (2003) Functional divergence prediction from evolutionary analysis: a case study of vertebrate hemoglobin. *Mol Biol Evol* 20, 1754-1759

43.   Seoighe, C*., et al.* (2003) Significantly different patterns of amino acid replacement after gene duplication as compared to after speciation. *Mol Biol Evol* 20, 484-490

44.   Conant, G.C*., et al.* (2007) Modeling amino acid substitution patterns in orthologous and paralogous genes. *Mol Phylogenet Evol* 42, 298-307

45.   Levasseur, A*., et al.* (2006) Tracking the connection between evolutionary and functional shifts using the fungal lipase/feruloyl esterase A family. *BMC Evol Biol* 6, 92

46.   Tennessen, J.A. (2008) Positive selection drives a correlation between non-synonymous/synonymous divergence and functional divergence. *Bioinformatics* 24, 1421-1425
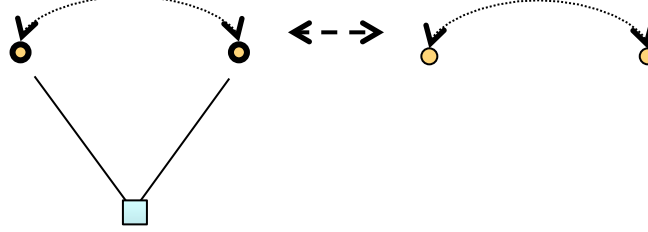
47. Lynch, V.J. (2007) Inventing an arsenal: adaptive evolution and neofunctionalization of snake venom phospholipase A2 genes. *BMC Evol Biol* 7, 2

48. Studer, R.A*., et al.* (2008) Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome Res* 18, 1393-1402

49. Clark, A.G*., et al.* (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature* 450, 203-218

50. Eyre-Walker, A. (2006) The genomic rate of adaptive evolution. *Trends Ecol Evol* 21, 569-575

51. Kosiol, C*., et al.* (2008) Patterns of positive selection in six Mammalian genomes. *PLoS Genet* 4, e1000144

52. Andolfatto, P. (2007) Hitchhiking effects of recurrent beneficial amino acid substitutions in the Drosophila melanogaster genome. *Genome Res* 17, 1755-1762

53. Davis, J.C., and Petrov, D.A. (2004) Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol* 2, E55

54. He, X., and Zhang, J. (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169, 1157-1164

55. Presser, A*., et al.* (2008) The evolutionary dynamics of the Saccharomyces cerevisiae protein interaction network after duplication. *Proc Natl Acad Sci U S A* 105, 950-954

56. Johnston, C.R*., et al.* (2007) Evaluation of whether accelerated protein evolution in chordates has occurred before, after, or simultaneously with gene duplication. *Mol Biol Evol* 24, 315-323

57. Davis, J.C., and Petrov, D.A. (2005) Do disparate mechanisms of duplication add similar genes to the genome? *Trends Genet* 21, 548-551

58. Cusack, B.P., and Wolfe, K.H. (2007) Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. *Mol Biol Evol* 24, 679-686

59. Ravi, V., and Venkatesh, B. (2008) Rapidly evolving fish genomes and teleost diversity. *Curr Opin Genet Dev* 18, 544-550

60. Liao, B.-Y., and Zhang, J. (2008) Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci U S A*, 0800387105

61. Altenhoff, A.M., and Dessimoz, C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol* 5, e1000262

62. Jordan, I.K*., et al.* (2004) Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol Biol* 4, 22
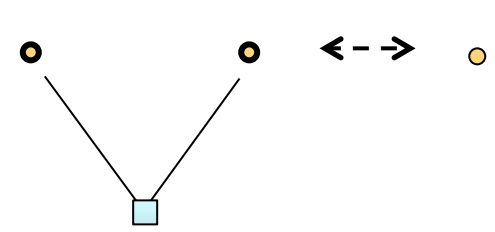
Common ancestor

Speciation 1

Duplication 1

Speciation 2

Duplication 2

A   Bα   Cα   Bβ   Cβ$_1$ Cβ$_2$
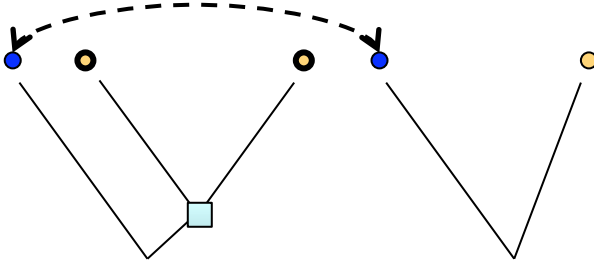
I. Divergence between duplicates

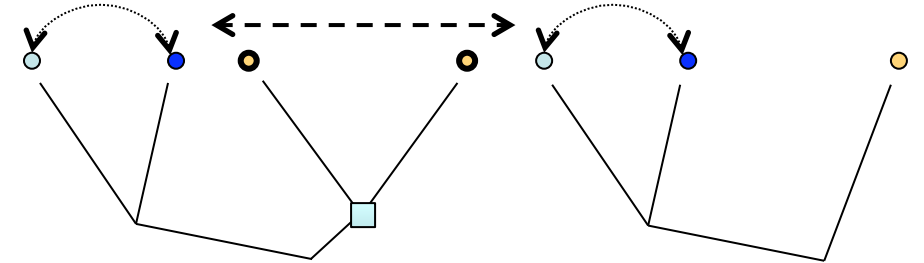II. Duplicate pairs *vs.* random pairs
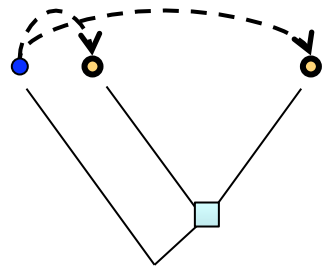
III. Duplicates *vs.* singletons

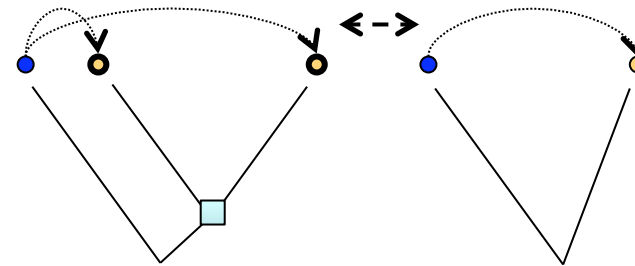IV. Orthologs of duplicates *vs.* orthologs of singletons

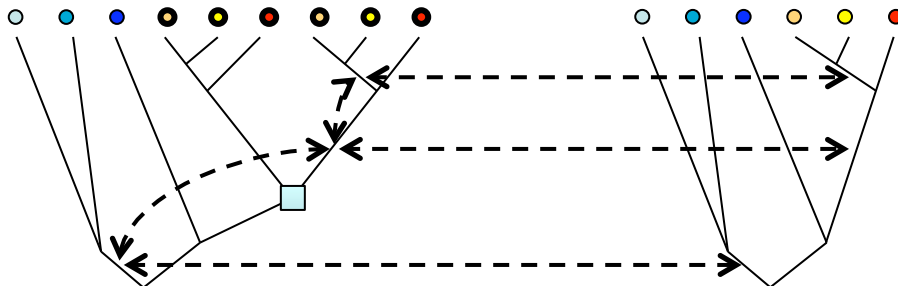V. Divergence between orthologs of duplicates *vs.* between orthologs of singletons

VI. Divergence relative to outgroup

VII. Divergence relative to outgroup of duplicates *vs.* singletons

VIII. Comparison of several singletons and duplicates per gene tree (sequences)

IX. Comparison of several singletons and duplicates per gene tree (functional data)

all to all