Increased Genome Sampling Reveals a Dynamic Relationship between Gene Duplicability and the Structure of the Primate Protein–Protein Interaction Network

Aoife Doherty,[†] David Alvarez-Ponce,[†] and James O. McInerney*

Department of Biology, National University of Ireland Maynooth, Maynooth, County Kildare, Ireland

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: james.o.mcinerney@nuim.ie.

Associate Editor: Michael Purugganan

Research article

Abstract

Although gene duplications occur at a higher rate, only a small fraction of these are retained. The position of a gene's encoded product in the protein–protein interaction network has recently emerged as a determining factor of gene duplicability. However, the direction of the relationship between network centrality and duplicability is not universal: In *Escherichia coli*, yeast, fly, and worm, duplicated genes more often act at the periphery of the network, whereas in humans, such genes tend to occupy the most central positions. Herein, we have inferred duplication events that took place in the different branches of the primate phylogeny. In agreement with previous observations, we found that duplications generally affected the most central network genes, which is presumably the process that has most influenced the trend in humans. However, the opposite trend—that is, duplication being more common in genes whose encoded products are peripheral in the network—is observed for three recent branches, including, quite counterintuitively, the external branch leading to humans. This indicates a shift in the relationship between centrality and duplicability during primate evolution. Furthermore, we found that genes encoding interacting proteins exhibit phylogenetic tree topologies that are more similar than expected for random pairs and that genes duplicated in a given branch of the phylogeny tend to interact with those that duplicated in the same lineage. These results indicate that duplication of a gene increases the likelihood of duplication of its interacting partners. Our observations indicate that the structure of the primate protein–protein interaction network affects gene duplicability in previously unrecognized ways.

Key words: gene duplicability, protein–protein interaction network, network evolution, primate evolution.

Introduction

One of the key insights provided by fully sequenced genomes is the pervasiveness of gene duplication and loss in all organisms (Ohno 1970; Zhang 2003), which has resulted in modern-day genomes being replete with multigene families and a confusing pattern of orthologs and paralogs distributed throughout life on the planet. However, genes widely differ in their propensity to retain duplicates, whereas some gene families are represented by dozens or even hundreds of members in a given genome, others remain as singleton genes over time. This observation naturally leads to the key question concerning why genes duplicate and the even bigger question of what constraints exist that might prevent duplications from occurring or at least retard their rate of occurrence.

Gene duplication is often the key for understanding the origin and evolution of important advantageous traits. For example, the acquisition of color vision in vertebrates is the result of the duplication of retinal visual pigment genes (Yokoyama 2002), and salivary amylase gene copy number is positively correlated with dietary starch intake in human populations (Perry et al. 2007). On the other hand, gene duplication is a significant factor in the pathogenesis of various diseases such as cancer (Slamon et al. 1987; Lahortiga et al. 2007). A duplicated gene is very likely to be lost unless it offers a selective advantage to the organism in which it is found, and therefore, only a fraction of duplicated genes are retained after duplication (Ohno 1970; Lipinski et al. 2011). Over the past decade, the combination of genomic and functional data has allowed us to identify the factors correlating with gene duplicability, that is, the tendency to retain both gene copies after duplication. These factors include gene function (Marland et al. 2004) and complexity (Papp et al. 2003; Yang et al. 2003; He and Zhang 2005), subcellular location (Prachumwat and Li 2006), and timing of expression during development (Castillo-Davis and Hartl 2002; Yang and Li 2004). Yet, a large fraction of the variability of gene duplicability remains unexplained.

Genes and proteins rarely act in isolation, and over the past few years, in particular, we have been gaining a better understanding of the complex networks of interactions in which these molecules find themselves. The high throughput accumulation of interactomic data now allows us to investigate the relationship between the patterns of molecular evolution of genes and the position that their encoded products occupy in protein–protein interaction networks (PINs) (see Cork and

Purugganan 2004; Eanes 2011; Zera 2011; Alvarez-Ponce et al. 2012). The position of a protein in the network can be measured from its network centrality, which can be computed as its degree (number of proteins with which it interacts), betweenness (number of shortest paths between protein pairs to which it belongs), or closeness (the inverse of the average distance to all other proteins in the network) (Borgatti 2005; Mason and Verwoerd 2007). Some aspects of the evolution of genes have been shown to be affected by the centrality of their encoded products in the PIN (e.g., Luisi et al. 2012). For instance, genes occupying the most central positions tend to be more selectively constrained (Fraser et al. 2002; Hahn and Kern 2005; Lemos et al. 2005). Although gene duplicability is also affected by centrality, the direction of the relationship between centrality and duplicability is not universal. In *E. coli*, yeast, and fly, singleton genes tend to occupy more central positions in the network than duplicated genes (Hughes and Friedman 2005; Prachumwat and Li 2006; Makino et al. 2009). A possible explanation for this phenomenon is that duplication of a gene may disrupt the dosage balance of the interactions in which it is involved (Veitia 2002; Papp et al. 2003), and this may have more deleterious effects for the most highly connected genes. Conversely, duplicated genes tend to be more central than singleton genes in the human PIN (Liang and Li 2007), which is a derived character resulting from the high duplicability of metazoan-specific genes (D'Antonio and Ciccarelli 2011). However, it remains unclear why this different pattern is observed in humans. These contrasting observations indicate that, although network position has a clear effect on a gene's duplicability, the relationship between duplicability and PIN centrality has undergone modification in the vertebrate lineage. This dynamic behavior of the relationship between centrality and duplicability opens the question of whether more shifts have taken place during evolution and, if so, how often did they occur and when.

Further evidence for the dependence between the position of genes in an interaction network and their patterns of evolution comes from the observation that genes encoding interacting proteins tend to exhibit correlated evolutionary histories (for a review, see Lovell and Robertson 2010). For example, their rates of evolution are more similar than expected from random protein pairs (Fraser et al. 2002; Lemos et al. 2005; Alvarez-Ponce et al. 2011). This similarity is generally attributed to molecular co-evolution or to interacting proteins being subject to similar evolutionary forces and it can be potentially used to infer protein–protein interactions from sequence data (Codoñer and Fares 2008; Fares et al. 2011). For instance, several studies have shown that interacting genes manifest phylogenetic histories that are more similar than expected in a random network, as evidenced by the similarity in the lengths of the branches in the phylogeny. However, gene tree similarities have usually been assessed using the mirrortree approach, which relies on the underlying distance matrices (Goh et al. 2000; Pazos and Valencia 2001; Pazos et al. 2008). It is less clear whether the actual phylogenetic trees inferred from interacting proteins are more topologically similar than expected from random protein pairs. In fact, Kelly and Stumpf (2010) found only negligible evidence for such an increased level of similarity between pairs of trees inferred from interacting proteins in sets of yeast orthologous sequences. However, both the mirrortree approach and the approach used by Kelly and Stumpf rely on sets of 1:1 orthologs. Although computationally convenient, this approach does not address the potential gene tree similarity resulting from similar duplication histories. Almost 20 years ago, it was hypothesized that interacting genes may tend to exhibit topologically similar phylogenetic trees owing to co-duplication at similar evolutionary times (Fryxell 1996). Arguably, duplication of a gene with interacting partners may be deleterious unless the interacting genes co-duplicate soon after or before the event (Papp et al. 2003). Alternatively, the functional diversification of duplicated genes could be facilitated by a pre-existing heterogeneity in proteins that interact with their products (Fryxell 1996). Although a number of examples of correlated tree topologies for interacting genes have been reported (e.g., Fryxell 1996; Koretke et al. 2000; Alvarez-Ponce et al. 2009), an analysis at the level of the entire interactome has not been carried out to date.

Herein, we combine comparative genomics and protein–protein interaction data to explore the relationship between the structure of the primate PIN and the duplicability of genes encoding its components. For that purpose, we inferred the gene duplication events that took place in each of the branches of a phylogeny consisting of six primates and one rodent and evaluated the dependence between the duplicability of genes and the position of their encoded products in the PIN. The results revealed a complex relationship between network position and duplicability. We found that 1) in agreement with previous observations, duplicated genes act at the most central positions of the human PIN; however, when we examined the trend across different portions of the primate phylogeny, the opposite (i.e., gene duplication preferentially affecting genes whose encoded products are peripheral in the PIN) was observed for genes duplicated in the external branch leading to humans and the two internal branches subtending the human/chimpanzee and the human/chimpanzee/gorilla clades, indicating that the relationship between duplicability and centrality has undergone modification more than once during animal evolution; 2) genes encoding interacting proteins exhibit more similar tree topologies than expected in a random network; and 3) genes that duplicated in a given branch tend to interact with genes that duplicated in the same branch, indicating that the duplication of a gene increases the likelihood of duplication of its interacting partners in the network. Taken together, these results indicate that the structure of the primate network constrains the patterns of duplication of their components at multiple levels and in a dynamic manner.

Materials and Methods

Genomic Data

We retrieved all protein-coding sequences (CDSs) and family assignments for human, chimpanzee, gorilla, orangutan, macaque, marmoset, and mouse from the Ensembl database (version 61; Flicek et al. 2011). We eliminated the following from our analyses: 1) coding sequences that were interrupted by a stop codon or whose length was not a multiple of three; 2) sequences that had not been assigned to any gene family; and 3) gene families consisting of less than four sequences. After this filtering, a dataset comprising 125,999 genes belonging to 12,158 gene families was retained.

Phylogenetic Tree Reconstruction and Duplication Inference

For each gene family, we aligned the protein sequences using MUSCLE (Edgar 2004). The resulting protein alignments were used to guide the alignment of the corresponding CDSs using TranslatorX (Abascal et al. 2010). The CDS alignments were subsequently used to reconstruct Bayesian phylogenetic trees using SPIMAP (Rasmussen and Kellis 2011). Gene duplications were inferred using the species/gene tree reconciliation approach implemented in the SPIMAP software and the species overlap method (Huerta-Cepas et al. 2007; Gabaldón 2008) implemented in the ETE package (Huerta-Cepas et al. 2010). For these analyses, we used the reference species tree provided in the study by Benton et al. (2009) (fig. 1).

We assigned each duplication event to a branch of the reference species tree and to one or more human genes. For each duplication node, we examined the species represented in the descendant leaves. We assigned the duplication event to the branch preceding the deepest node in the reference species tree whose descendants include all the species affected by the duplication. For instance, if sequences descending from a duplication node included sequences from all great apes, the duplication event was assigned to the branch subtending the radiation of the great apes. Subsequently, we assigned the duplication event to the set of human genes that are the result of this duplication event or are the closest human homologs to the genes involved in this duplication. If there was at least one human gene in the set of descendant leaves of a duplication node, the duplication event was assigned to this human gene or set of genes. Otherwise, we systematically examined the parental node of that node until the descendant leaves contained at least one human homolog.

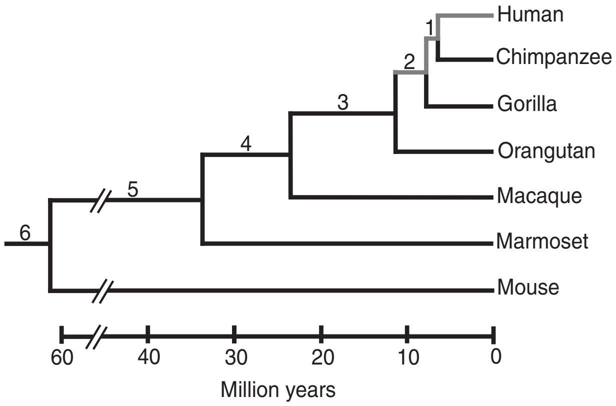

Fig. 1. Phylogeny of the species included in the analysis. Divergence dates were retrieved from the study by Benton et al. (2009). The number above each internal branch (1–6) is the name that we have assigned to that branch. Branches for which duplicated genes tend to be less connected than nonduplicated genes are represented in gray.

Network-Level Analysis

The human interactome was assembled from the interactions available from the BioGRID database version 3.1.81 (Stark et al. 2011). Only nonredundant physical interactions among pairs of human proteins with an Ensembl ID were considered. The network (termed PIN0) contains 9,087 proteins connected by 39,883 interactions. For each protein, degree was computed as the number of interacting partners, and betweenness and closeness centralities were computed using the NetworkX package (http://networkx.lanl.gov/). Proteins not represented in the PIN0 network were not used in network-level analyses.

We evaluated whether the phylogenetic trees of genes encoding interacting proteins were more similar than expected in a random network and whether genes that duplicated in a given branch of the species tree tend to interact with genes that duplicated in the same branch. For that purpose, a subnetwork containing only proteins with a nontrivial tree was used (PIN1; supplementary fig. S1, Supplementary Material online). We used as statistics the average tree topological similarity of interacting proteins (see below) and the number of interactions among proteins encoded by duplicated genes. The statistical significance of measured network parameters was evaluated from an ensemble of 250 or 10,000 randomized networks. Random networks were generated using a network rewiring approach. Each random network was generated from PIN1 by repeatedly choosing two edges at random (e.g., A–B and C–D) and swapping them (yielding A–D and C–B, or A–C and B–D). This operation was iterated 100 × m times on each random network, where m is the number of edges. Therefore, each random network contains the same nodes, the same number of edges, and the same degree for each node as the original network. P values were computed as the proportion of random networks with a parameter value higher or equal to the observed one.

To discard the potential impact of confounding network features in our results, analyses were repeated on two subnetworks of PIN1. PIN2 is a subnetwork of PIN1 with no self-interactions or interactions among proteins encoded by paralogous genes; and PIN3 is a subnetwork of PIN2 without interactions among proteins encoded by genes locating in the same chromosome arm (supplementary fig. S1, Supplementary Material online). A separate network ensemble was generated for each of these networks. The same restrictions used to generate each subnetwork were imposed to the corresponding ensembles, only allowing edge swaps respecting these restrictions.

Comparison of Phylogenetic Trees

We used the "tree comparison" program from the treeKO package (Marcet-Houben and Gabaldón 2011) to compare the tree topologies of pairs of interacting proteins. The "strict distance" was used. Trees were rooted in the branch that minimized the number of gene duplications in the tree.

Age of Human Genes

To establish the age of each human gene, we carried out a similarity search against the nr database (downloaded on 12 October 2010; Pruitt et al. 2007) using the BLASTP algorithm (Altschul et al. 1990). Only genes that aligned to more than 80% of the query sequence were retained. If at least 5% of the hits corresponded to nonmetazoan genomes, the human gene was considered to be of premetazoan ancestry (i.e., "ancient").

Results

Identifying Duplication Events in the Primate Phylogeny

We retrieved all CDSs for six primates (human, chimpanzee, gorilla, orangutan, macaque, and marmoset) and one rodent (mouse). After filtering the dataset (see Materials and Methods), we retained a total of 125,999 genes belonging to 12,158 gene families (supplementary table S1, Supplementary Material online). For each family, we reconstructed a phylogenetic tree using a Bayesian approach. Using these phylogenetic trees, we inferred the duplication and loss events that took place during the evolution of each family using the gene tree/species tree reconciliation approach (Goodman et al. 1979; Page 1994). This algorithm compares each gene tree with an established species tree topology (fig. 1), and discrepancies between the two are attributed to duplication or loss events. Because inference of gene losses is methodologically problematic (Hahn 2007), only gene duplications are considered in the current analysis. In addition to the gene tree/species tree reconciliation approach, we used the reconciliation-independent species overlap method (Huerta-Cepas et al. 2007; Gabaldón 2008), which is based on the pattern of species overlap in the descendant leaves of each duplication node. The gene tree/species tree reconciliation approach inferred a total of 22,969 duplications across the studied phylogeny, whereas the more conservative species overlap method inferred 15,814 duplications (table 1 and supplementary table S2, Supplementary Material online). Unless otherwise stated, the results reported throughout this article correspond to duplications inferred using the gene tree/species tree reconciliation method; however, we carried out all analyses in parallel using both approaches, with qualitatively equivalent results. These results are detailed in the relevant tables, and all analyses and data are available in the Supplementary Material online or on request from the authors.

We estimated an overall gene duplication rate of 0.00348 duplications/gene/My across the phylogeny of the studied species. However, we found that the duplication rate varied widely across the different branches of the tree, ranging from 0.0012 duplications/gene/My on the chimpanzee external branch to 0.0252 duplications/gene/My on the internal branch subtending the human, chimpanzee, and gorilla clade (labeled as branch 2; see table 1). This represents a greater than 20-fold difference in duplication rate between these branches. The remarkable acceleration in the rate of gene duplication in branch 2 has been described previously and has been suggested to be the result of changes in the effective population size or the generation time during the evolution of the great apes (Marques-Bonet et al. 2009). In agreement with previous reports (Hahn et al. 2007), we observed an increased rate of gene duplication in the primate lineage (0.00388 duplications/gene/My) compared with the mouse branch (0.0018 duplications/gene/My). Furthermore, we observed an increased rate of duplication in the great apes (0.0041 duplications/gene/My) compared with the average rate in primates, also consistent with previous observations (Fortna et al. 2004; Hahn et al. 2007).

Of particular interest in the assessment of gene duplication is the issue of what kinds of genes have duplicated and in which evolutionary time. For each branch in the species tree, we obtained a list of human genes that are either the result of

Table 1. Summary Statistics for Each Branch of the Studied Phylogeny (Species/Gene Tree Reconciliation Method).

Branch Name	Branch Length (My)	Number of Duplications	Rate of Duplication	Number of Human Homologs	Ancient Human Homologs (%)	Ancient Human Homologs in PIN0 (%)
Human	6.5	495	0.0037	790	9.49	18.95
Chimpanzee	6.5	157	0.0012	217	10.60	30.77
Gorilla	8.0	424	0.0025	426	22.06	38.79
Orangutan	11.2	292	0.0013	342	22.22	41.82
Macaque	23.5	902	0.0018	731	32.15	45.41
Marmoset	33.7	1,526	0.0021	1,043	30.97	39.99
Mouse	61.5	2,584	0.0018	796	12.81	28.38
Branch 1	1.5	90	0.0030	179	17.32	26.47
Branch 2	3.2	1,655	0.0252	1,805	23.16	29.13
Branch 3	12.3	1,770	0.0071	1,906	21.30	26.99
Branch 4	10.2	2,127	0.0099	2,274	23.04	31.25
Branch 5	27.8	3,220	0.0055	3,108	21.30	26.15
Branch 6	—	7,727	—	8,668	20.72	22.83

duplication events that occurred in that branch or the closest human homologs to the genes involved in the duplications that occurred at that branch (see Materials and Methods). We considered whether each of the resulting gene lists was enriched in certain Gene Ontology (GO; Ashburner et al. 2000) terms. For that purpose, we compared the frequency of each GO term in the list of duplicated genes with the rest of the human genome using the FatiGO software (Al-Shahrour et al. 2004), which specifically seeks to find significant associations between GO terms and lists of genes. A total of 67 unique biological processes are enriched among genes duplicated in any of the external branches of the phylogeny (supplementary table S3, Supplementary Material online). In general, we observed enrichment in the "reproduction," "transcription," "translation," and "environment perception" GO categories, in agreement with previous results (e.g., Demuth and Hahn 2009; Huerta-Cepas et al. 2007). Interestingly, we observed a clear enrichment in GO categories associated with olfactory transduction in mouse-specific duplications, as reported previously by Niimura and Nei (2005, 2007).

From these results, we can conclude that our dataset and treatments of the data are in line with previous work. In this study, we have conducted an interactome-wide analysis of gene duplication.

The Relationship between Centrality and Duplicability Underwent Modification during Primate Evolution

Having identified the genes that underwent duplication in each branch of the phylogeny of the studied species, we sought to investigate the relationship between the structure of the network and the duplicability of its components. For that purpose, we assembled a human interactome (termed PIN0) from all interactions available in the BioGRID database (Stark et al. 2011). For each gene in the network, we computed three centrality measures (degree, betweenness, and closeness) and compared their values for nonduplicated genes and genes that underwent duplication in any branch of the phylogeny. In agreement with previous observations in the human interactome (Liang and Li 2007; D'Antonio and Ciccarelli 2011), we found that duplicated genes occupy more central positions in the human PIN than nonduplicated genes using the Mann–Whitney U test ($P = 2.89 \times 10^{-13}$ for degree; $P = 3.01 \times 10^{-10}$ for betweenness; and $P = 2.11 \times 10^{-14}$ for closeness; supplementary table S4, Supplementary Material online, and fig. 2). Crucially, however, this analysis only takes into account whether genes underwent duplication in any branch of the phylogeny, and therefore, it does not consider the specific branches on the species tree in which those duplications occurred. A more interesting analysis is to consider duplications that happened at approximately the same time and whether different parts of the interactome were perturbed by duplication at different times.

We conducted an analysis that partitioned duplication events into the branches in the phylogeny in which they occurred. We observed that duplicated genes exhibit a higher average degree (i.e., number of interacting partners) than nonduplicated genes in 10 of the 13 branches of the species tree, with statistically significant differences in 5 of the branches (supplementary table S4, Supplementary Material online, and fig. 2). Unexpectedly, the opposite trend (i.e., a higher degree for nonduplicated genes) is observed in the three remaining branches (the external branch leading to the human lineage and internal branches 1 and 2), with statistically significant differences in two of these branches (the human branch and internal branch 1; supplementary table S4, Supplementary Material online, and fig. 2). We obtained similar results when we used betweenness and closeness as the measures of network centrality and when the more conservative species overlap method was used as the method for inferring duplication events (supplementary table S4, Supplementary Material online). Therefore, despite the general tendency of duplications to occur at the most central genes of the network, the relationship between centrality and duplicability has inverted during the primate radiation.

These observations presents us with a picture of the relationship between network position and gene duplicability

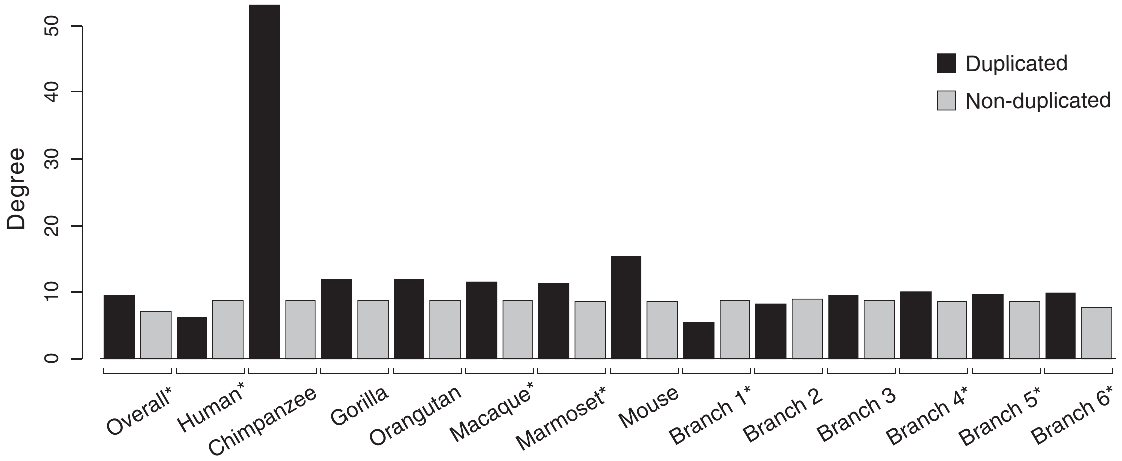

Fig. 2. Degree for proteins encoded by duplicated and nonduplicated genes in each branch of the phylogeny. Results for the corresponding statistical tests are reported in supplementary table S4, Supplementary Material online. The asterisk indicates statistically significant differences.

that is more complex than has been reported previously and that up to now was assumed to be the general rule for vertebrates. To gain a more complete understanding into the relationship between the structure of the network and the patterns of duplication of its components, we considered whether duplications of genes encoding interacting proteins were correlated.

Interacting Proteins in the Human PIN Tend to Exhibit Topologically Similar Phylogenetic Trees

For each pair of interacting proteins, we compared the topologies of the corresponding phylogenetic trees using the treeKO algorithm (Marcet-Houben and Gabaldón 2011). We used the "strict distance," which takes into account both the patterns of speciation and the duplication and loss patterns. For this analysis, we used a subnetwork of PIN0 (termed PIN1; see supplementary fig. S1, Supplementary Material online) that contained only proteins encoded by genes capable of reconstructing nontrivial phylogenetic trees (those belonging to gene families with four or more members). According to the treeKO algorithm, trees derived from interacting proteins exhibit an average distance of $D = 0.319$. To assess the significance of this value, we compared it with a null distribution obtained from a set of randomized networks with the same nodes, number of interactions, and degree for each node as the original network (PIN1; see Materials and Methods). Of 250 randomized versions of PIN1, none showed a D value lower than or equal to the observed one (average value for the simulations, $D = 0.339$; $P < 0.004$; fig. 3), indicating that interacting proteins in the human interactome manifest tree topologies that are more similar than expected from a random network.

This similarity might be the result of molecular coevolution of genes encoding interacting proteins. However, a number of features of PINs might also produce such a similarity, and their effects should be ameliorated as much as possible to eliminate potential sources of confounding bias. First, PINs are known to be enriched in self-interactions (i.e., interactions among identical proteins) and interactions among proteins encoded by paralogous genes (Ispolatov et al. 2005; Pereira-Leal et al. 2007; Alvarez-Ponce and McInerney 2011). Because genes involved in these interactions are represented in the same phylogenetic trees, this enrichment could potentially contribute to the low observed D value. To discard this possibility, analyses were repeated in a subnetwork of PIN1 in which all such interactions were removed (PIN2; see supplementary fig. S1, Supplementary Material online). We found that interacting proteins still exhibit a higher similarity than expected in a random network ($D = 0.331$; average value for the simulations, $D = 0.338$; $P < 0.004$), indicating that these features do not affect our observations. Second, duplication events sometimes affect large chromosomal regions, thereby involving simultaneous duplication of multiple adjacent genes, which would consequently have similar duplication histories. In addition, genes encoding interacting proteins tend to cluster together in the genome (Lee and Sonnhammer 2003; Makino and McLysaght 2008). Taken together, these tendencies may also contribute to the similarity in tree topologies observed among interacting proteins. However, the topological similarity of trees in the observed interactome is still significantly higher than expected at random when interactions involving proteins encoded by genes that localize to the same chromosome arm are also removed ($D = 0.331$; average value for the simulations, $D = 0.338$; $P < 0.004$ for PIN3; supplementary fig. S1, Supplementary Material online).

These results indicate that genes encoding interacting proteins manifest more similar tree topologies than expected from random pairs and that this pattern is independent of the enrichment of the network in self-interactions, interactions among paralogous genes, and interactions among genes that co-localize in the genome. This similarity can potentially be the result of genes that encode interacting proteins exhibiting similar duplication histories. To test this possibility, we investigated whether the duplications of interacting genes tend to occur in the same branches of the species tree.

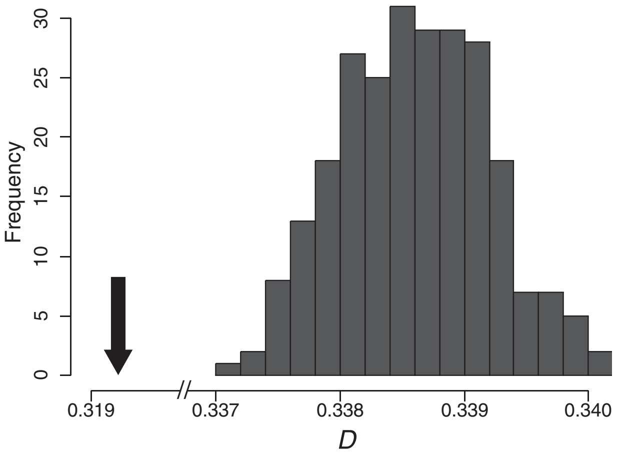

FIG. 3. Topological distance between the trees corresponding to pairs of interacting proteins in the human interactome. The observed value in the actual interactome (PIN1) is represented as an arrow, and the distribution inferred from 250 randomized networks is represented as a histogram.

Genes Encoding Interacting Proteins Tend to Co-duplicate in the Same Branches of the Phylogeny

We considered whether the human interactome was enriched in interactions among proteins encoded by duplicated genes. For that purpose, we computed the number of interactions involving genes that have undergone duplication in any branch of the phylogeny ($N = 22{,}988$ in PIN1) and compared this number to the null distribution obtained from a collection of 10,000 random networks. None of these random networks exhibits an N value higher than or equal to the observed one ($P < 0.0001$), indicating that duplicated genes tend to interact with each other in the real network. This result holds when self-interactions and interactions among paralogs ($N = 21{,}872$; $P < 0.0001$ for PIN2), and interactions between genes locating in the same chromosome arm ($N = 21,152$; $P < 0.0001$ for PIN3), are removed from the analyses.

We carried out an equivalent analysis for genes duplicated in each of the 13 branches of the studied phylogeny; that is, we examined whether genes that duplicated in a given branch tend to interact with genes that duplicated in the same branch. For each branch *i*, we computed the number of interactions between genes that underwent duplication in that branch, N_i, and evaluated its statistical significance as above. When all interactions are considered (PIN1), the N_i values are significantly higher than expected from a random network in all 13 branches ($P < 0.05$; supplementary table S5, Supplementary Material online), indicating that genes that have undergone duplication in each of these branches tend to interact with each other. When self-interactions and interactions among paralogs are removed (PIN2), the N_i values are higher than the average values for the random networks for 10 of the 13 branches, with statistically significant differences in 4 of the branches (the external branches leading to gorilla, marmoset, and mouse, and internal branch 6; supplementary table S5 and supplementary fig. S2, Supplementary Material online). Qualitatively equivalent results were obtained when interactions among genes in the same chromosome arm were also removed from the analysis (PIN3; supplementary table S5, Supplementary Material online). Similar results are obtained using the species overlap method (supplementary table S5, Supplementary Material online). These results indicate that although the tendency of genes that duplicated in a given branch to interact with each other is in part the result of the enrichment of the network in self-interactions and interactions among paralogs (Ispolatov et al. 2005; Pereira-Leal et al. 2007; Alvarez-Ponce and McInerney 2011), these features cannot completely account for the observed trend.

Discussion

We used phylogenetic methods to accurately determine the branches of the primate phylogeny at which each gene family duplicated and investigated the relationship between a gene's pattern of duplication and the position of its encoded product in the primate PIN. We addressed this dependency from three perspectives. First, we evaluated the relationship between network centrality of a protein and the duplicability of the encoding gene in the different branches of the studied phylogeny. Second, we tested whether interacting proteins manifest topologically similar phylogenetic trees, in particular, when we look beyond the analysis of 1:1 orthologs. Finally, we considered whether genes encoding interacting proteins tend to duplicate at the same branches of the phylogeny. In all three cases, we found new significant results, with some patterns being more complex than previously thought.

The Dynamic Relationship between Centrality and Duplicability

In *E. coli*, yeast, and fly, genes occupying central positions tend to remain singleton, whereas those acting at the periphery of the network can more often retain duplicated copies (Hughes and Friedman 2005; Prachumwat and Li 2006; Makino et al. 2009). This has been attributed to the deleterious effects of altering the dosage balance of protein–protein interactions (Veitia 2002; Papp et al. 2003). In contrast with the pattern observed in the aforementioned organisms, duplicated genes tend to be more central in the human interactome (Liang and Li 2007; D'Antonio and Ciccarelli 2011), indicating that the relationship between duplicability and centrality has undergone modification during animal evolution. The pattern observed in the human interactome has been attributed to the possibility that the involvement of a gene in a higher number of interactions would facilitate the functional diversification of paralogs, for example, through tissue specialization or that highly connected proteins would be required in higher dosages (Liang and Li 2007).

Consistent with previous observations in the human genome (Liang and Li 2007; D'Antonio and Ciccarelli 2011), we found that primate genes that duplicated in any branch of the species tree tend to be more central than singleton genes (supplementary table S4, Supplementary Material online, and fig. 2). According to the dosage balance hypothesis, duplication of a gene would be deleterious unless its interacting partners underwent co-duplication soon after or before (Papp et al. 2003). An extreme example of co-duplication is whole genome duplication (WGD), which maintains the relative dosage of all balanced sets (Veitia 2004, 2005). Therefore, the high content of ohnologs (i.e., genes resulting from the two WGD events that occurred in early vertebrate evolution; Wolfe 2000) in mammalian genomes (Nakatani et al. 2007; Makino and McLysaght 2010) might potentially provide an explanation for the lack of a negative association between duplicability and centrality in mammals. Indeed, when the relationship between duplicability and centrality was analyzed separately for genes duplicated in each branch of the phylogeny, we found that genes that duplicated in the ancestral branch to all studied species (branch 6; fig. 1), which include ohnologs, tend to be more central than genes that did not duplicate at that branch (supplementary table S4, Supplementary Material online, and fig. 2). However, we also observed the same pattern in most of the other branches of the phylogeny (all of them post-WGD): Duplicated genes encode more central proteins than nonduplicated genes (supplementary table S4, Supplementary Material online, and fig. 2). This indicates that the preferential duplication of central genes is an ongoing process that can be observed in relatively recent branches (e.g., the macaque branch, which encompasses the last ~23.5 My; Benton et al. 2009; figs. 1 and 2) and not solely the result of WGD.

Unexpectedly, the opposite relationship between duplicability and centrality is observed in the external branch leading to humans and in the internal branches subtending the human/chimpanzee (branch 1) and the human/chimpanzee/gorilla (branch 2) clades, with statistically significant differences for the human branch and branch 1 (fig. 1). That is to say, in contrast to the overall trend, genes that duplicated in these lineages tend to occupy more peripheral positions in the network than nonduplicated genes (supplementary table S4, Supplementary Material online, and fig. 2), resembling the pattern observed in *E. coli*, yeast, and fly (Hughes and Friedman 2005; Prachumwat and Li 2006; Makino et al. 2009). Therefore, the relationship between duplicability and centrality seems to have undergone a reversal during the evolution of great apes, revealing that this relationship is highly dynamic.

D'Antonio and Ciccarelli (2011) recently showed that the particular relationship between duplicability and centrality observed in humans is the result of the high content of the human genome in genes that arose late in evolution. Human genes of ancient (premetazoan) origin exhibit the same pattern as observed in *E. coli*, yeast, and fly (duplicated genes are less central), whereas human genes of more recent origin (those that originated within the metazoans) exhibit the opposite trend (duplicated genes tend to be more central). This contrasting pattern observed among ancient and new human genes could potentially provide an explanation for the different relationship between centrality and duplicability that we observe in the different branches of the phylogeny if duplications in the human branch and internal branch 1 involved preferentially ancient genes. However, we found that the proportion of ancient genes among genes that duplicated in these branches (9.49–17.32%) is generally lower than for genes that duplicated in the other branches of the phylogeny (table 1) (qualitatively similar results are obtained when the analysis is restricted to genes represented in the human interactome; table 1), indicating that the different age of genes that duplicated in the different branches of the phylogeny is not the factor responsible for the heterogeneity in the relationship between duplicability and centrality observed here.

Genes Encoding Interacting Proteins Exhibit Correlated Tree Topologies and Duplication Histories

To gain further insight into the relationship between the structure of the primate PIN and the duplicability of its components, we then considered whether genes encoding interacting proteins exhibit tree topologies that are more similar than expected from a random pair of proteins. We found that interacting genes exhibit phylogenetic trees with a higher similarity than expected from a random PIN (fig. 3). This tendency is not the result of the enrichment of the human interactome in self-interactions and interactions among paralogs (Ispolatov et al. 2005; Pereira-Leal et al. 2007; Alvarez-Ponce and McInerney 2011) or the clustering in the genome of genes encoding interacting proteins (Lee and Sonnhammer 2003; Makino and McLysaght 2008). Our observations contrast with those by Kelly and Stumpf (2010). They found only negligible evidence for pairs of yeast interacting proteins presenting phylogenetic trees topologically more similar than random pairs of proteins. At least three possible reasons might account for the different results obtained in both studies. First, they analyzed the yeast interactome, whereas we focused on primates; therefore, it might be possible that both interactomes would exhibit a different trend. Second, the datasets used by Kelly and Stumpf (2,528–5,109 proteins and 5,728–21,283 interactions) were remarkably smaller than the one used here, which could have limited statistical power in their analyses. Finally, they inferred phylogenetic trees from 1:1 orthologous sets, whereas we used entire gene families. Although computationally convenient, using 1:1 orthologous sets removes the effect of duplication and loss events in the tree topologies. Therefore, the different results obtained in the analysis by Kelly and Stumpf (2010) and our analysis may also potentially be the result, at least partially, of interacting genes exhibiting similar patterns of duplication and/or loss.

We found that the number of interactions between genes that underwent duplication at any branch of the phylogeny is higher than expected from a random network. This observation indicates that duplicated human genes tend to interact with each other in the PIN (supplementary fig. S2, Supplementary Material online), lending support to the hypothesis that duplication of a gene may increase the likelihood of duplication of its interacting partners (Fryxell 1996). This trend holds true when genes that duplicated in each particular branch of the phylogeny are analyzed separately (supplementary table S5 and supplementary fig. S2, Supplementary Material online). Although the significance vanishes for most of the branches when self-interactions and interactions among paralogs are removed, the trend remains significant for four of the branches (the external branches leading to gorilla, marmoset, and mouse and internal branch 6; supplementary table S5 and supplementary fig. S2, Supplementary Material online). Interestingly, these branches include the three longest branches in the phylogeny (the external branches leading to mouse and marmoset and internal branch 6; see fig. 1), suggesting that perhaps the lack of significance in the remaining branches may be the result of reduced statistical power in the shorter branches. Alternatively, the absence of significance in these branches might be a consequence of the reduced efficiency of selective mechanisms favoring the co-duplication of interacting genes in the same branches of the phylogeny. Indeed, we might expect that the selective advantage of duplicating the interacting partner of a protein would be often small. Furthermore, changes in gene dosage can be compensated by mechanisms different from complementary gene duplication, such as changes in expression levels, or may even be accommodated by stochastic variation in levels of protein expression. Therefore, the tendency of interacting genes to co-duplicate in the same branches of the species tree may be observed in organisms only in which natural selection is highly efficient. Primates have a lower effective population size than rodents, which is thought to involve a reduced efficiency of natural selection (Ohta 1973; Lynch 2007); therefore, evolutionary forces promoting the co-duplication of genes encoding interacting proteins may be less efficient in primates.

Conclusion

Taken together, our analyses indicate that the position of proteins in the primate PIN has an effect on the patterns of duplication of their encoding genes, indicating that the network imposes constraints on the fate of genes encoding its components. First, gene duplicability depends on the centrality of the encoded products in the network, although, interestingly, the relationship between centrality and duplicability has varied during primate evolution. Second, interacting proteins exhibit similar duplication histories, tending to co-duplicate in the same branches of the phylogeny. Furthermore, we observed that interacting genes exhibit topologically similar phylogenetic trees, possibly owing to these correlated duplication histories.

Although separate analysis of individual genomes represents a valuable tool to provide a first glance at the patterns of gene duplication, this approach only allows a binary classification of genes as singleton or duplicated, thus providing only an aggregate overview. A more comprehensive characterization of duplication events can be gained by including multiple genomes in the analysis. This comparative approach allows, for instance, assigning duplication events to particular branches in the species tree. When applied to a relatively small selection of mammals, this approach allowed us to observe a dynamical relationship between the structure of the PIN and the patterns of duplication of genes encoding its components. Future work is warranted to understand how the structure of the PIN has influenced gene duplicability in other lineages. In particular, it will be interesting to see the relationship between duplicability and network position in organisms with effective population sizes that are larger than those for mammals. In addition, we note that a limitation of our analysis is that currently available interactomic data are highly incomplete and subject to a high rate of false negatives (Bader et al. 2004; Deeds et al. 2006). The future availability of more complete and accurate interactomes will allow a deeper understanding of the relationship between duplicability and centrality.

Supplementary Material

Supplementary tables S1–S5 and figs. S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Toni Gabaldón for helpful comments on the manuscript. The authors would like to thank the Irish Centre for High End Computing and the NUI Maynooth HPC facility. This work was supported in part by a Science Foundation Ireland grant 09/RFP/EOB2510 to J.O.M. and an Irish Research Council for Science Engineering and Technology grant to A.D.

References

Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* 38:W7.

Al-Shahrour F, Díaz-Uriarte R, Dopazo J. 2004. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 20:578.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215:403–410.

Alvarez-Ponce D, Aguadé M, Rozas J. 2009. Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 *Drosophila* genomes. *Genome Res.* 19:234–242.

Alvarez-Ponce D, Aguadé M, Rozas J. 2011. Comparative genomics of the vertebrate insulin/TOR signal transduction pathway: a network-level analysis of selective pressures. *Genome Biol Evol.* 3:87–101.

Alvarez-Ponce D, Guirao-Rico S, Orengo DJ, Segarra C, Rozas J, Aguade M. 2012. Molecular population genetics of the insulin/TOR signal transduction pathway: a network-level analysis in *Drosophila melanogaster*. *Mol Biol Evol.* 29:123–132.

Alvarez-Ponce D, McInerney JO. 2011. The human genome retains relics of its prokaryotic ancestry: human genes of archaebacterial and eubacterial origin exhibit remarkable differences. *Genome Biol Evol.* 3:782–790.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT. 2000. Gene ontology: tool for the unification of biology. *Nat Genet.* 25:25.

Bader JS, Chaudhuri A, Rothberg JM, Chant J. 2004. Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol.* 22:78–85.

Benton MJ, Donoghue PCJ, Asher RJ. 2009. Calibrating and constraining molecular clocks. In: Hedges SB, Kumar S, editors. The timetree of life. Oxford: Oxford University Press. p. 35–86.

Borgatti SP. 2005. Centrality and network flow. *Soc Network.* 27:55–71.

Castillo-Davis CI, Hartl DL. 2002. Genome evolution and developmental constraint in *Caenorhabditis elegans*. *Mol Biol Evol.* 19:728–735.

Codoñer FM, Fares MA. 2008. Why should we care about molecular coevolution? *Evol Bioinform Online.* 4:29–38.

Cork JM, Purugganan MD. 2004. The evolution of molecular genetic pathways and networks. *Bioessays* 26:479–484.

D'Antonio M, Ciccarelli FD. 2011. Modification of gene duplicability during the evolution of protein interaction network. *PLoS Comput Biol.* 7:e1002029.

Deeds EJ, Ashenberg O, Shakhnovich EI. 2006. A simple physical model for scaling in protein-protein interaction networks. *Proc Natl Acad Sci U S A.* 103:311–316.

Demuth DP, Hahn MW. 2009. The life and death of gene families. *Bioessays* 31:29–39.

Eanes WF. 2011. Molecular population genetics and selection in the glycolytic pathway. *J Exp Biol.* 214:165–171.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792.

Fares MA, Ruiz-Gonzalez MX, Labrador JP. 2011. Protein coadaptation and the design of novel approaches to identify protein-protein interactions. *IUBMB Life.* 63:264–271.

Flicek P, Amode MR, Barrell D, et al. (52 co-authors). 2011. Ensembl 2011. *Nucleic Acids Res.* 39:D800–D806.

Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T. 2004. Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol.* 2:e207.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. *Science* 296:750–752.

Fryxell KJ. 1996. The coevolution of gene family trees. *Trends Genet.* 12:364–369.

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

Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. 2000. Co-evolution of proteins with their interaction partners. *J Mol Biol.* 299:283–293.

Goodman M, Czelusniak J, Moore GW, Romero-Herrera A, Matsuda G. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst Biol.* 28:132–163.

Hahn MW. 2007. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol.* 8:R141.

Hahn MW, Demuth JP, Han SG. 2007. Accelerated rate of gene gain and loss in primates. *Genetics* 177:1941.

Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol.* 22:803–806.

He X, Zhang J. 2005. Gene complexity and gene duplicability. *Current Biol.* 15:1016–1021.

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

Huerta-Cepas J, Dopazo J, Gabaldón T. 2010. ETE: a python environment for tree exploration. *BMC Bioinformatics* 11:24.

Hughes AL, Friedman R. 2005. Gene duplication and the properties of biological networks. *J Mol Evol.* 61:758–764.

Ispolatov I, Yuryev A, Mazo I, Maslov S. 2005. Binding properties and evolution of homodimers in protein-protein interaction networks. *Nucleic Acids Res.* 33:3629–3635.

Kelly W, Stumpf M. 2010. Trees on networks: resolving statistical patterns of phylogenetic similarities among interacting proteins. *BMC Bioinformatics.* 11:470.

Koretke KK, Lupas AN, Warren PV, Rosenberg M, Brown JR. 2000. Evolution of two-component signal transduction. *Mol Biol Evol.* 17:1956–1970.

Lahortiga I, De Keersmaecker K, Van Vlierberghe P, Graux C, Cauwelier B, Lambert F, Mentens N, Beverloo HB, Pieters R, Speleman F. 2007. Duplication of the MYB oncogene in T cell acute lymphoblastic leukemia. *Nat Genet.* 39:593–595.

Lee JM, Sonnhammer EL. 2003. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.* 13:875–882.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol Biol Evol.* 22:1345–1354.

Liang H, Li WH. 2007. Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends Genet.* 23:375–378.

Lipinski KJ, Farslow JC, Fitzpatrick KA, Lynch M, Katju V, Bergthorsson U. 2011. High spontaneous rate of gene duplication in *Caenorhabditis elegans*. *Current Biol.* 21:306–310.

Lovell SC, Robertson DL. 2010. An integrated view of molecular coevolution in protein-protein interactions. *Mol Biol Evol.* 27:2567–2575.

Luisi P, Alvarez-Ponce D, Dall'olio GM, Sikora M, Bertranpetit J, Laayouni H. 2012. Network-level and population genetics analysis of the insulin/TOR signal transduction pathway across human populations. *Mol Biol Evol.* 29:1379–1392.

Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.

Makino T, Hokamp K, McLysaght A. 2009. The complex relationship of gene duplication and essentiality. *Trends Genet.* 25:152–155.

Makino T, McLysaght A. 2008. Interacting gene clusters and the evolution of the vertebrate immune system. *Mol Biol Evol.* 25:1855–1862.

Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A.* 107:9270–9274.

Marcet-Houben M, Gabaldón T. 2011. TreeKO: a duplication-aware algorithm for the comparison of phylogenetic trees. *Nucleic Acids Res.* 39:e66.

Marland E, Prachumwat A, Maltsev N, Gu Z, Li WH. 2004. Higher gene duplicabilities for metabolic proteins than for nonmetabolic proteins in yeast and *E. coli*. *J Mol Evol.* 59:806–814.

Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LDW, Jiang Z, Baker C, Malfavon-Borja R, Fulton LA. 2009. A burst of segmental duplications in the african great ape ancestor. *Nature* 457:877.

Mason O, Verwoerd M. 2007. Graph theory and networks in biology. *IET Syst Biol.* 1:89–119.

Nakatani Y, Takeda H, Kohara Y, Morishita S. 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.* 17:1254–1265.

Niimura Y, Nei M. 2005. Comparative evolutionary analysis of olfactory receptor gene clusters between humans and mice. *Gene* 346:13–21.

Niimura Y, Nei M. 2007. Extensive gains and losses of olfactory receptor genes in mammalian evolution. *PLoS One.* 2:e708.

Ohno S. 1970. Evolution by gene duplication. Berlin: Springer-Verlag.

Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* 246:96–98.

Page RDM. 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst Biol.* 43:58.

Papp B, Pal C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424:194–197.

Pazos F, Juan D, Izarzugaza JM, Leon E, Valencia A. 2008. Prediction of protein interaction based on similarity of phylogenetic trees. *Methods Mol Biol.* 484:523–535.

Pazos F, Valencia A. 2001. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.* 14:609–614.

Pereira-Leal JB, Levy ED, Kamp C, Teichmann SA. 2007. Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol.* 8:R51.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet.* 39: 1256–1260.

Prachumwat A, Li WH. 2006. Protein function, connectivity, and duplicability in yeast. *Mol Biol Evol.* 23:30–39.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35:D61–D65.

Rasmussen MD, Kellis M. 2011. A bayesian approach for fast and accurate gene tree reconstruction. *Mol Biol Evol.* 28:273.

Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. 1987. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* 235:177.

Stark C, Breitkreutz BJ, Chatr-Aryamontri A, et al. (15 co-authors). 2011. The BioGRID interaction database: 2011 update. *Nucleic Acids Res.* 39:D698–D704.

Veitia RA. 2002. Exploring the etiology of haploinsufficiency. *Bioessays* 24:175–184.

Veitia RA. 2004. Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. *Genetics* 168: 569–574.

Veitia RA. 2005. Paralogs in polyploids: one for all and all for one? *Plant Cell.* 17:4–11.

Wolfe K. 2000. Robustness—it's not where you think it is. *Nat Genet.* 25: 3–4.

Yang J, Li WH. 2004. Developmental constraint on gene duplicability in fruit flies and nematodes. *Gene* 340:237–240.

Yang J, Lusk R, Li WH. 2003. Organismal complexity, protein complexity, and gene duplicability. *Proc Natl Acad Sci U S A.* 100:15661.

Yokoyama S. 2002. Molecular evolution of color vision in vertebrates. *Gene.* 300:69–78.

Zera AJ. 2011. Microevolution of intermediary metabolism: evolutionary genetics meets metabolic biochemistry. *J Exp Biol.* 214:179–190.

Zhang J. 2003. Evolution by gene duplication: an update. *Tr Ecol Evol.* 18: 292–298.