Parallel evolution of metazoan mitochondrial proteins

Galya V. Klink[1] and Georgii A. Bazykin[1,2,3,4,*]

[1]Institute for Information Transmission Problems (Kharkevich Institute) of the Russian Academy of Sciences, Moscow 127051, Russia

[2]Pirogov Russian National Research Medical University, Moscow 117997, Russia

[3]Faculty of Bioengineering and Bioinformatics and Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow 119234, Russia

[4] Skolkovo Institute of Science and Technology, Skolkovo, 143025, Russia

Corresponding author: E-mail: gbazykin@iitp.ru

Abstract

Amino acid propensities at amino acid sites change with time due to epistatic interactions or changing environment, affecting the probabilities of fixation of different amino acids. Such changes should lead to an increased rate of homoplasies (reversals, parallelisms, and convergences) at closely related species. Here, we reconstruct the phylogeny of twelve mitochondrial proteins from several thousand metazoan species, and measure the phylogenetic distances between branches at which either different alleles originated due to divergent substitutions, or the same allele originated repeatedly due to homoplasies. The mean phylogenetic distance between parallel substitutions is ~20% lower than the mean phylogenetic distance between divergent substitutions, indicating that a variant fixed in a species is more likely to be deleterious in more phylogenetically remote species, compared to a more closely related species. These findings are robust to artefacts of phylogenetic reconstruction or of pooling of sites from different conservation classes or functional groups, and show that single-position fitness landscapes change at rates similar to rates of amino acid changes.

Key words: fitness landscape, epistasis, parallel substitutions, heteropecilly, mitochondria

Introduction

Amino acid preferences at a site, or single-position fitness landscape (SPFL, Bazykin 2015), change in the course of evolution, so that a variant conferring high fitness in one species may confer low fitness in another, either due to changes at interacting genomic sites or in the environment. These changes can be observed through phylogenetic patterns, in particular, through a non-uniform distribution of amino acid substitutions giving rise to a particular variant (homoplasies) along the phylogeny. Indeed, when a certain amino acid repeatedly arises at a particular site in a certain phylogenetic clade, but is never observed at this site in another clade, this implies that the relative fitness conferred by this variant in the former clade is higher. Different types of homoplasies – reversals, parallelisms and convergencies – have been found to be clustered on the phylogenies (Rogozin et al. 2008; Povolotskaya and Kondrashov 2010; Naumenko et al. 2012; Goldstein et al. 2015; Zou and Zhang 2015), and an attempt has been made to estimate the rate at which SPFLs change from such data (Usmanova et al. 2015).

SPFL changes in metazoan mitochondrial proteins were previously inferred from amino acid usage patterns (Breen et al. 2012), but this approach has been criticized as sensitive to the underlying assumptions regarding fitness distributions (McCandlish et al. 2013). Here, we develop an approach for the study of phylogenetic clustering of homoplasies at individual protein sites, and apply it to deep alignments of mitochondrial proteins of metazoans (Breen et al. 2012) together with their reconstructed phylogenies. Our approach compares the distributions of distances between parallel and divergent substitutions to infer robustly changes in relative fitness of different variants at a site between branches of the phylogenetic tree.

Materials and Methods

Alignment and phylogeny

We obtained multiple-species alignments of 12 mitochondrial proteins of metazoans from (Breen et al. 2012), and analyzed alignment columns with fewer than 1% gaps (which comprised 77% of all sites). As there is no accepted phylogenetic tree for this large and diverse set of species spanning a wide range of phylogenetic distances, we took a hybrid approach to reconstructing their phylogeny. First, we constrained the tree topology using the curated taxonomy-based phylogeny of the ITOL (Interactive tree of life) project (Letunic and Bork 2007). By requiring the presence of each species in the ITOL database, we were left with >900 metazoan species for each protein (Table 1). The resulting topology was not fully resolved, and contained multifurcations. We then used RAxML 8.0.0 (Stamatakis 2014) under the GTR-Γ model to resolve multifurcations and to estimate the branch lengths. Finally, ancestral states were reconstructed using codeml program of the PAML package (Yang 1997) under the substitution matrix and the value of the parameter alpha of the gamma distribution inferred by RAxML.

Clustering of substitutions on a phylogeny

Using the inferred states of amino acid sites at each node, we inferred the positions of all substitutions at all protein sites on the phylogenies of the corresponding proteins. For each ancestral amino acid at a site, we defined parallel substitutions as those giving rise to the same derived amino acid, and divergent substitutions, as those giving rise to different derived amino acids (fig. 1A). We considered only those pairs of substitutions that happened in phylogenetically independent branches, i.e., such that one was not ancestral to the other. For subsequent analysis, we analyzed only the set of homoplasy-informative sites, i.e., sites that have at least one pair of parallel substitutions and one pair of divergent substitutions from the same ancestral amino acid. The phylogenetic distance between a pair of substitutions was defined as the distance (measured in the number of amino acid substitutions per amino acid site inferred by RAxML) between the centers of the edges where those substitutions have occurred, i.e., the sum of the distances from the centers of these edges to the last common ancestor of the two substitutions (fig. 1A).

To compare the distances between parallel and divergent substitutions while circumventing the potential biases associated with pooling sites and amino acids with different properties (see below), we subsampled the pairs of parallel and divergent substitutions, and analyzed the distances in this subset. For this, for each ancestral amino acid at homoplasy-informative sites, we picked randomly min(N_{paral}, N_{diverg}) pairs of parallel substitutions and the same number of pairs of divergent substitutions, where N_{paral} and N_{diverg} are the numbers of parallel and divergent substitutions originating from this amino acid. We repeated this procedure for all ancestral amino acids at all sites, thus obtaining two equal-sized subsamples of parallel and divergent substitutions, and measured all distances in these resulting subsamples. The parallel to divergent ratio (P/D) for each 0.1 distance bin was calculated by dividing the number of parallel pairs of substitutions by the number of divergent pairs of substitutions that had occurred at a distance from each other falling into this bin (main text), or not exceeding the given distance (Supplementary Figures). These two statistics are closely related respectively to the O-ring and Ripley's K statistics widely used in spatial ecology to measure aggregation in communities (Wiegand and A. Moloney 2004).

To obtain mean values and confidence intervals of each statistic, we bootstrapped sites in 1000 replicates, each time repeating the entire resampling procedure.

Robustness of tree shape

For each branch of the phylogenetic tree, we obtained the bootstrap support value in 100 bootstrap replicates using RAxML, and performed our analyses only with pairs of substitutions such that the nodes ancestral to both substitutions had bootstrap support higher than 90, ensuring the robustness of these nodes.

Comparison of vertebrate and invertebrate substitution matrices

To infer distinct substitution matrices for vertebrates and invertebrates, we separately reconstructed the ITOL-constrained phylogenies of 980 vertebrate and 92 invertebrate species with RAxML under the GTR-Γ model using the concatenated sequence of ATP6 and ND1 genes. We compared the empirical GTR matrices reconstructed by RAxML.

Simulated evolution

For each gene, we simulated amino acid evolution using the evolver program of the PAML package (Yang 1997) using the phylogenetic tree, substitution matrix and alpha-parameter of the gamma-distribution output by RAxML for the corresponding gene. From the simulated amino acids at the leaves of the tree, we then reconstructed the ancestral states using codeml (PAML) under the same parameters.

Results

Phylogenetic clustering as evidence for SPFL changes

We devised an approach for analysis of the clustering of parallel substitutions at a site, which is robust to other specifics of the phylogenetic distribution of substitutions. Conceptually, our approach is similar to the one used previously (Povolotskaya and Kondrashov 2010; Goldstein et al. 2015; Zou and Zhang 2015). In addition, it is designed to control for any potential biases that can arise from pooling sites with different properties. Since it is difficult to obtain robust evidence for SPFL changes for an individual amino acid site even using large numbers of species, getting a significant signal of SPFL changes requires pooling different amino acid sites. The problem is that these sites may differ in their properties, and such differences may lead to artefactual evidence for SPFL changes, for the following reasons.

First, pooling of parallel and convergent substitutions giving rise to the same descendant variant, i.e., substitutions with the same and different ancestral variants, may provide artefactual evidence for SPFL changes due to reasons such as the structure of the genetic code. For example (fig. 1B), an amino acid A within a clade may arise repeatedly from the ancestral amino acid B_1 at a particular clade simply because the $B_1 \rightarrow A$ mutation is frequent. If another amino acid B_2 is more prevalent than B_1 at a different clade, and the $B_2 \rightarrow A$ mutation is less frequent, this will lead to an excess of substitutions giving rise to A in the former clade; this excess, however, is not an evidence for SPFL changes, but instead occurs for non-selective reasons. To control for this, we do not consider convergent substitutions, and separately consider the distributions of parallel and divergent substitutions from each ancestral variant B.

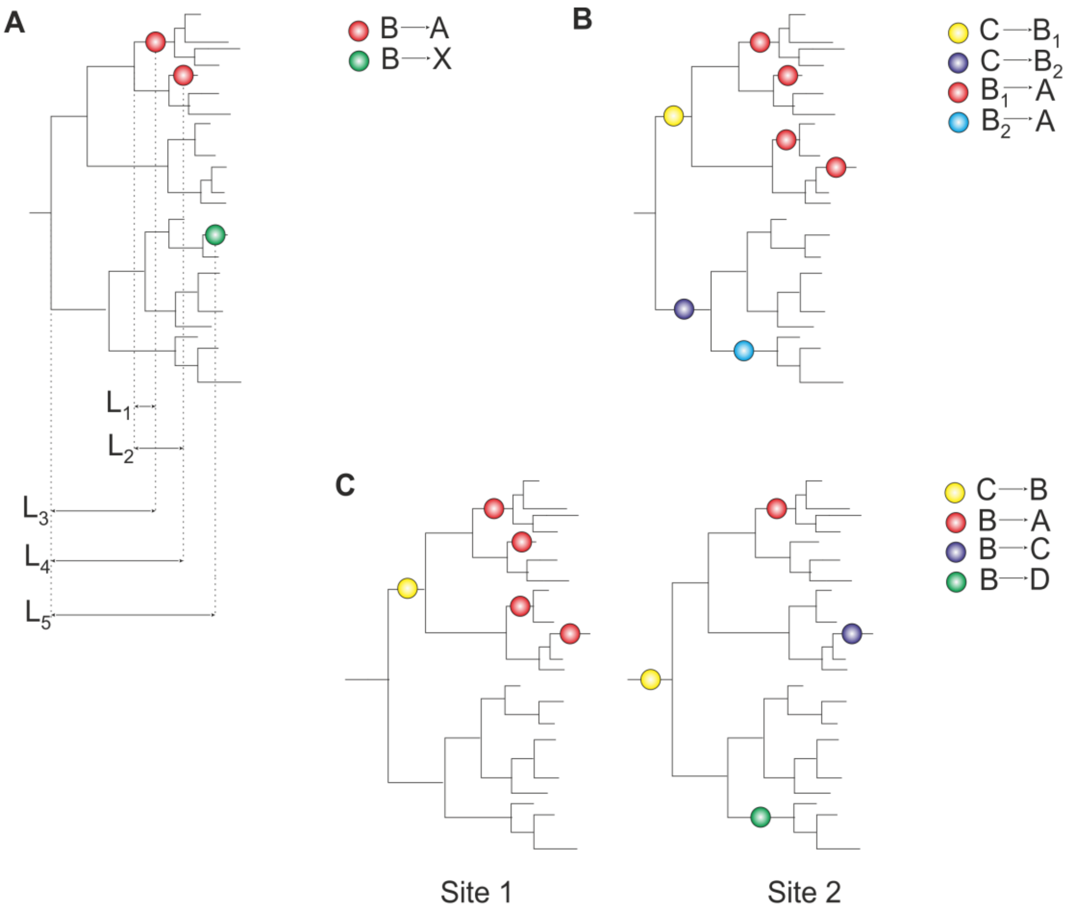


Figure 1. Inference of phylogenetic distances between parallel and divergent substitutions. Dots represent substitutions mapped to nodes of a phylogenetic tree. (A) For each pair of amino acids (B, A) at a particular amino acid site, we consider the distances between all parallel B→A substitutions (L_1+L_2), and distances between all divergent substitutions B→A and B→X (L_3+L_5, L_4+L_5), where X is any amino acid other than A and B. (B) The B_1→A substitution is more frequent than the B_2→A substitution, leading to an excess of homoplasies at small phylogenetic distances when parallel and convergent substitutions are pooled together. (C) Pooling of sites with different properties may also lead to an excess of homoplasies at small phylogenetic distances (see text).

Second, even independent consideration of different ancestral variants still permits clustering of homoplasies without SPFL changes when sites, and amino acids within sites, with diverse properties are pooled together. To illustrate this, assume that we analyze phylogenetic distances between parallel and divergent substitutions in a pooled sample of sites. Consider the hypothetical scenario in Figure 1C. At site 1, the amino acid B only resides within a relatively small clade. Therefore, both B→A and B→X substitutions are, by necessity, phylogenetically close to each other. By contrast, at site 2, the amino acid B is long living, and the distances between substitutions from it may be larger. If such sites also differ systematically in their amino acid propensities, this might lead to artefactual evidence for SPFL changes. For example, if sites where B is short-living (like site 1) also tend to be those where few amino acids confer high fitness (so that B→A substitutions are more frequent), while sites where B spans a large clade tend to be promiscuous with respect to the occupied amino acid (so that B→X substitutions are more frequent), pooling such sites may result in an excess of homoplasies within short phylogenetic distances.

We circumvent this problem by resampling matched sets of parallel and divergent substitutions. Specifically, at each homoplasy-informative site (see Methods), we consider different ancestral amino acids separately. For each ancestral amino acid B, we subsample our sets of pairs of substitutions: for each pair of parallel (B→A, B→A) substitutions, we randomly pick exactly one pair of divergent (B→A, B→X) substitutions from the same site. Finally, we pool together these subsamples from different sites, and analyze distances between parallel and divergent substitutions in this pooled set. This approach controls for any possible biases associated with differences in phylogenetic distributions of different ancestral amino acids. From the resulting subsets of distances, we calculate the ratio of the numbers of parallel to divergent substitutions (P/D) for each 0.1 distance window (see Methods).

Parallel substitutions in mitochondrial proteins are phylogenetically clustered

We applied this approach to the phylogenetic trees of 12 orthologous mitochondrial proteins of metazoans, each including >900 species. At the vast majority of sites, we observe many amino acid variants, in line with (Breen et al. 2012). By reconstructing the ancestral states and substitutions at each site, we observe that most of these variants have originated more than once, allowing us to study the phylogenetic distribution of homoplasies in detail (Table 1 and supplementary table 1, Supplementary Material online).

Table 1. Amino acid substitutions in mitochondrial genes of metazoans.

Gene	Species	Amino acid sites	Amino acids per site	Substitutions per site	Amino acids per site in simulation	Substitutions per site in simulation
ATP6	2931	186	9.5	152.1	12.1	154.63
COX1	4366	404	6.1	63.8	10.6	117.26
COX2	4131	165	8.9	137.2	12.4	206.83
COX3	2152	198	9.2	131.6	13.0	167.41
CYTB	5995	327	9.4	174.2	13.3	252.03
ND1	2013	253	8.7	92.9	12.7	124.42
ND2	5765	299	10.2	259.6	13.6	313.94
ND3	2766	94	9.7	182.3	12.8	194.37
ND4	2007	392	9.1	127.1	13.1	165.89
ND4L	1759	82	11.3	139.3	13.6	175.37
ND5	926	516	7.9	57.6	11.3	75.72
ND6	996	119	10.5	76.6	13.2	103.45

We observe an excess of parallel substitutions for species at small phylogenetic distances from each other (fig. 2-3), in line with the previous findings in vertebrates that used a smaller dataset (Goldstein et al. 2015). The P/D ratio is ~1.7 to 2.5 at phylogenetic distances less than 0.1, but drops to ~1 rapidly for larger distances (fig. 4 and supplementary fig. 1-2, Supplementary Material online). In simulated data, only a very weak decrease in the P/D ratio was observed which is possibly attributable to ancestral states reconstruction mistakes.

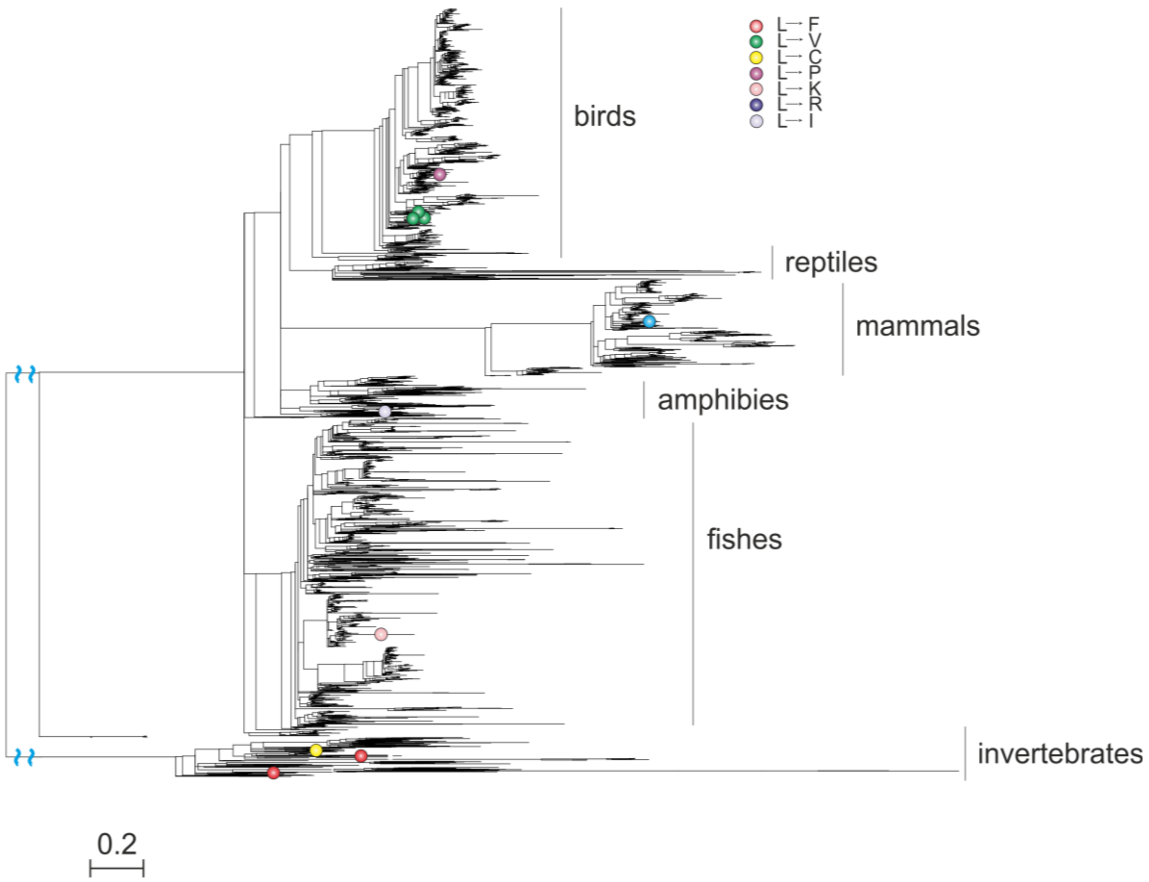


Figure 2. Parallel and divergent substitutions at site 202 of ATP6 (NCBI reference sequence numbering for the human sequence). The ancestral variant (L) has experienced multiple substitutions, which are scattered throughout the

phylogeny. However, the two parallel L→F substitutions occur in closely related species; the same is true for the three parallel L→V substitutions. Phylogenetic distances are in numbers of amino acid substitutions per site. The branches indicated with the blue waves are shortened by 1.2 distance units.

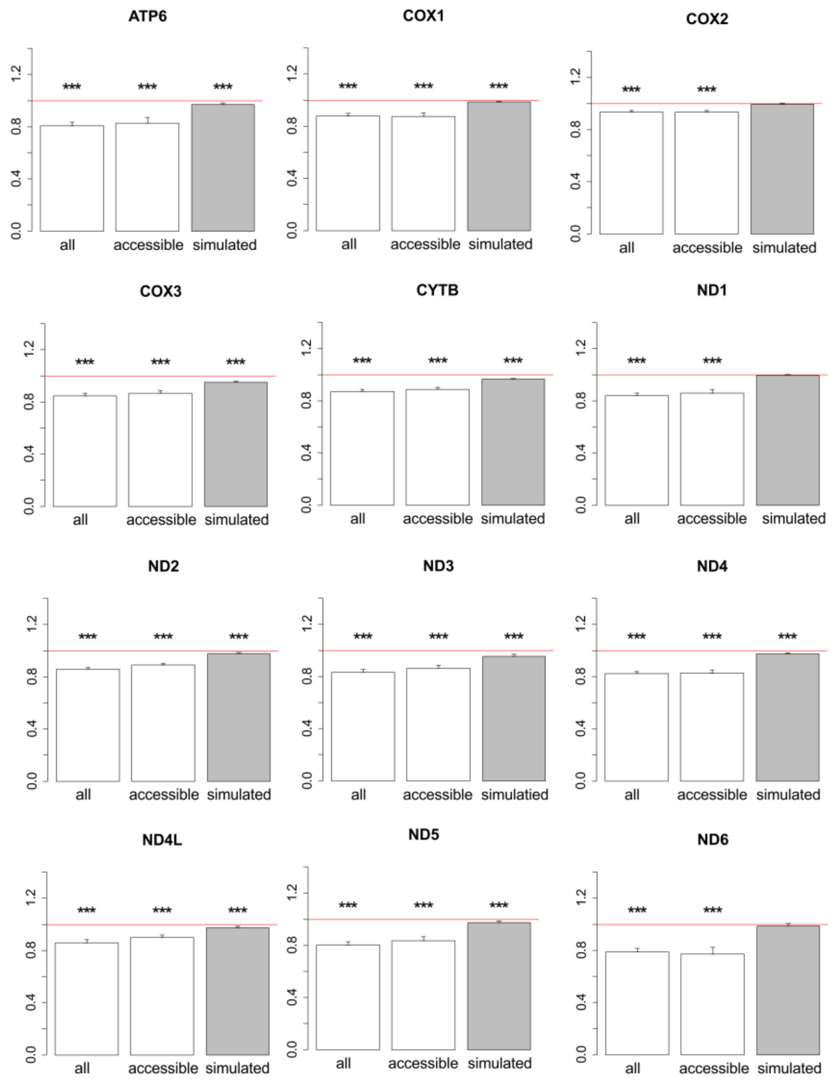


Figure 3. Ratios of phylogenetic distances between parallel and divergent substitutions in metazoan phylogenies. Values below 1 imply that the parallel substitutions are closer at the phylogeny to each other, compared to divergent substitutions. The bar height and the error bars represent respectively the median and the 95% confidence intervals obtained from 1,000 bootstrap replicates, and asterisks show the significance of difference from the one-to-one ratio (red line; *, P<0.05; **, P<0.01; ***, P<0.001). all, real data; accessible, real data only for substitutions from accessible amino acid pairs (see text); simulated, simulated data.

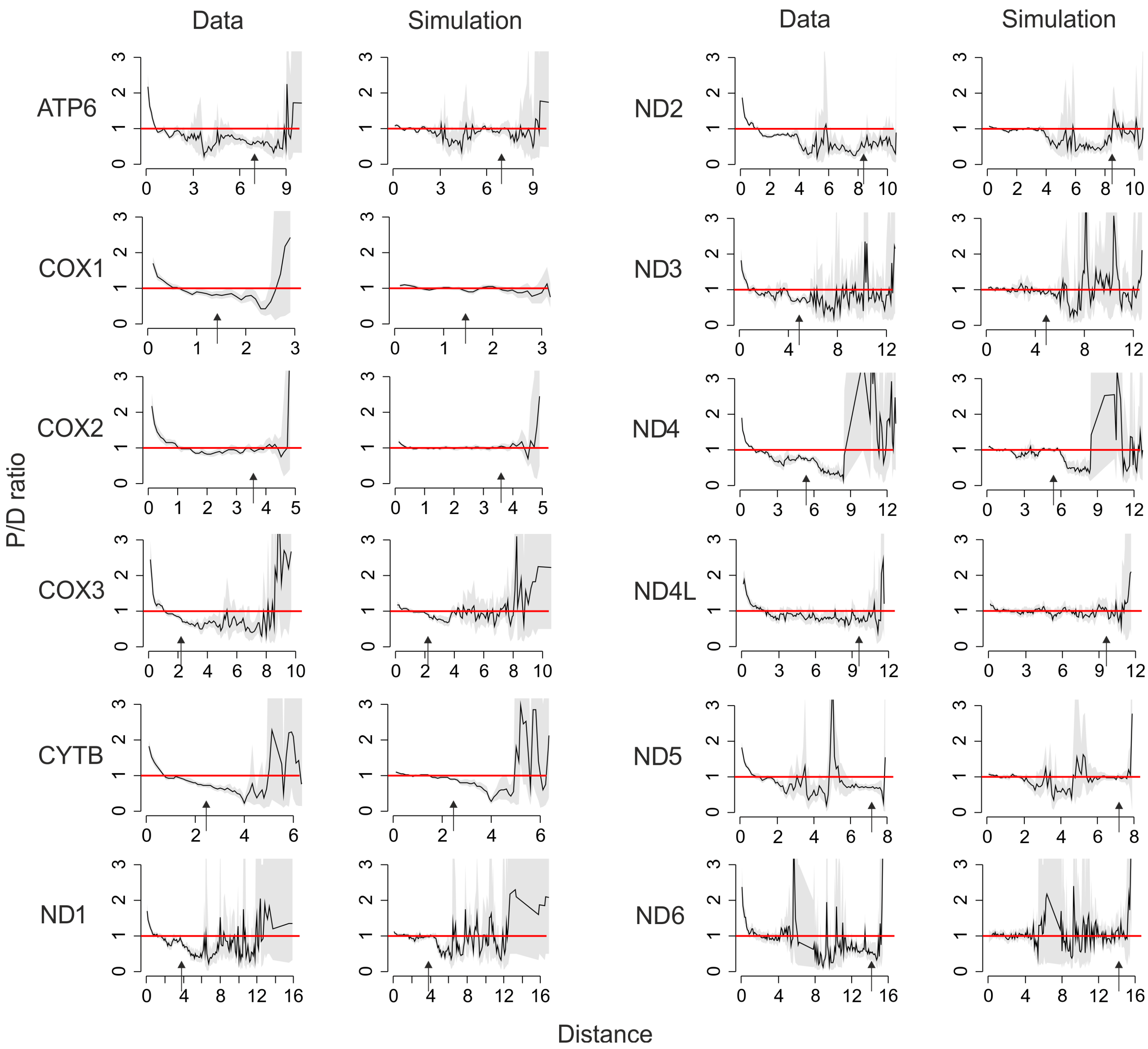


Figure 4. Higher fraction of parallel substitutions between closely related species. Horizontal axis, distance between branches carrying the substitutions, measured in numbers of amino acid substitutions per site (split into bins of 0.1). Vertical axis, P/D ratios for substitutions at this distance. Black line, mean; grey confidence band, 95% confidence interval obtained from 1000 bootstrapping replicates. The red line shows the expected P/D ratio of 1. Arrows represent the distance between human and *Drosophila.*

The rate at which the P/D ratio declines with phylogenetic distance varies a lot between genes. We asked whether this difference has to do with the intrinsic rate of protein evolution, which also varied strongly between genes. We used the number of substitutions between human and *Drosophila* as the proxy for the rate of protein evolution, with the higher values corresponding to rapidly evolving genes; and the phylogenetic distance at which the P/D ratio (which is initially always larger than one) reaches one, as the proxy for the rate of the decline of the P/D ratio, with the higher values corresponding to slower decline. Among the 10 analyzed genes for which sequences both for human and *Drosophila* were available, the P/D decline appeared to be somewhat faster in fast-evolving genes, although this trend was not significant (Spearman's test: R=0.53, p=0.11).

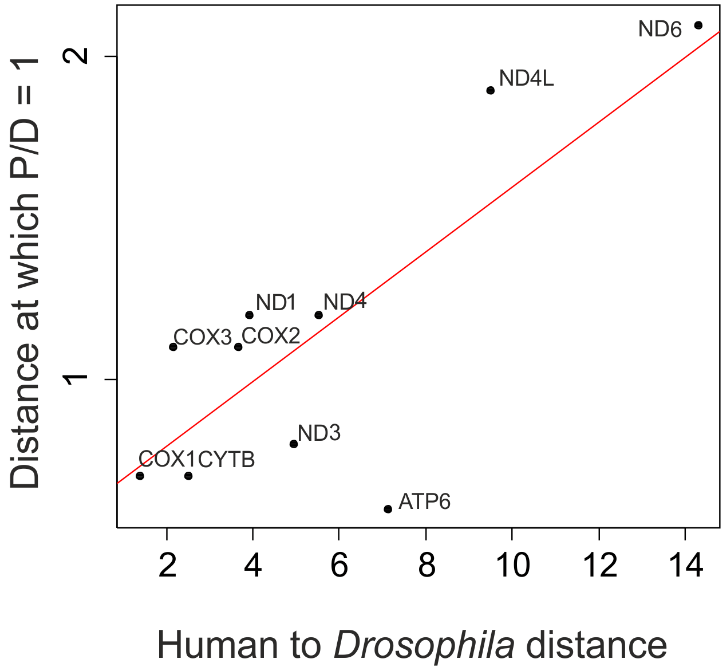


Figure 5. Faster decline of the P/D ratio for rapidly evolving genes. Horizontal axis, gene-specific phylogenetic distance between *Homo sapiens* and *Drosophila simulans*. Vertical axis, phylogenetic distance at which the P/D ratio reaches 1. Each dot represents one gene, line represents linear trend. Only ten genes for that *Drosophila simulans* sequence persists in a tree are taken.

Excess of parallel substitutions at small phylogenetic distances is not an artefact

Conceivably, the decline of the P/D ratio could be an artefact of erroneous phylogenetic reconstruction. Indeed, if a clade is erroneously split on a phylogeny, synapomorphies (shared derived character states) may be mistaken for parallel substitutions, and this is more likely for closely related species. As we used accepted species phylogenies, it is unlikely that such artefacts contribute to our analyses significantly. We tested their contribution by performing our analyses only with those pairs of substitutions where the phylogenetic nodes ancestral to both substitutions had bootstrap support higher than 90. This procedure was very conservative, because branches with parallel substitutions are expected to have a reduced bootstrap support, as such substitutions cause attraction of the branches where they occur in phylogenetic reconstruction. Indeed, this procedure removed the vast majority of parallel substitutions at very small phylogenetic distances, leading to P/D<1 at such distances. Still, at somewhat larger distances, P/D>1 was still observed in all genes (except a few where the difference from 1 was not significant;

supplementary fig. 3, Supplementary Material online). Therefore, the P/D ratio higher than 1 is not due to artefacts of phylogenetic reconstruction.

Changes in the P/D ratio with increasing phylogenetic distance imply changes in the rate of the B→A substitution relative to other substitutions. The rate of a substitution is the product of the mutation and fixation probabilities, and changes in the substitution rate may arise from differences in either of these processes between clades.

Can the changes in substitution rates with phylogenetic distance be explained by changes in mutation rates? There can be two scenarios for this. First, the mitochondrial mutational spectra could differ between clades, potentially leading to differences in the rate at which a particular mutation occurs. If the mutation rate corresponding to a particular substitution is much higher in a particular clade, compared to the rest of the phylogeny, this may lead to an excess of homoplasies falling into this clade. However, the differences between the mutation matrices for even the two most remote parts of the considered tree – vertebrates and invertebrates – are minor (supplementary fig. 4, Supplementary Material online). Most of the change in the P/D ratio occurs at very small phylogenetic distances (fig. 4), where the mutation matrices are very similar, and unlikely to contribute to our effect.

Second, even if the changes in the P/D ratio are not due to changes in the overall mutation matrix, they may still arise from differences in codon usage. This may be observed if amino acid B tends to be encoded by different codons in the two clades, and the rate of the parallel B→A substitution is higher in the clade where B is encoded by a codon that predisposes to this mutation. To test this, we defined “accessible” amino acid pairs as those (B, A) pairs where A can be reached through a single nucleotide substitution from any B codon, and considered such accessible pairs independently. In this subset, the excess of parallel changes at small phylogenetic distances was also observed (fig. 3), which means that it is not caused by the structure of the genetic code.

Discussion

The rate at which a specific substitution occurs is a monotonic function of fitness differences between the descendant and the ancestral variants, and changes in the substitution rates in the course of evolution indicate that these fitness differences, and thus the SPFL, change. Obtaining the entire substitution matrix for individual amino acid sites is problematic even given hundreds of species (Rodrigue 2013); still, the changes in the relative substitution rates can be inferred using some summary statistics. The change in the extent of parallelism in the course of evolution is one such convenient statistic: as a substitution becomes more deleterious, its rate decreases, and it becomes more sparsely distributed on the phylogeny.

We observe that parallel substitutions in the evolution of metazoan mitochondrial proteins are phylogenetically clustered; i.e., that such substitutions are more likely to occur in the phylogenetic vicinity of each other, compared with divergent substitutions. As a result, the distance between two parallel substitutions on the phylogenetic tree is, on average, ~20% lower than the distance between divergent substitutions, or than expected if their rate was constant across the tree (fig. 3). We show that these results are not artefacts of phylogenetic reconstruction or of pooling together sites and amino acids with different properties. Our results cannot be explained by a simple covarion model, in which a site alternates between neutral and constrained (Fitch and Markowitz 1970; Fitch 1971), as the changes we observe are not associated with changes in the overall substitution rates. For the same reason, they also cannot be explained by a broader class of heterotachy models in which the overall rates of evolution of a site vary with time (Lopez et al. 2002; Yang and Nielsen 2002; Murrell et al. 2012), but require heteropecilly (Tamuri et al. 2009; Roure and Philippe 2011), i.e., variation with time of rates of individual substitutions. We show that these differences are not explainable by systematic gene- or genome-wide differences in substitution matrices between clades, which may result from differences in mutation patterns, selection for nucleotide or amino acid usage, or gene conversion.

Instead, they reflect changes in single-position fitness landscapes (Bazykin 2015) that accumulate in the course of evolution. Indeed, site-specific differences in the rate of a substitution leading to a particular amino acid imply that the relative fitness of this amino acid relative to other amino acids at this site changes with time. Decrease in this frequency with phylogenetic distance may be caused by a decline in the fitness of this allele, and/or by an increase in the fitness of other alleles; it is hard to distinguish between these possibilities with the available data, although both factors likely play a role (Naumenko et al. 2012).

The numbers of substitutions to the same or to another amino acid, i.e. convergent and divergent substitutions at different phylogenetic distances, have been used previously to characterize evolution. In ancient proteins, the rate of convergence monotonically decreases with phylogenetic distance, and half of the reversals were estimated to become forbidden after 10% protein divergence (Povolotskaya and Kondrashov 2010). The ratio of the rates of convergent and divergent substitutions drops by more than twofold with phylogenetic distance within vertebrates (Goldstein et al. 2015). Similarly, the rate of convergence decreases with phylogenetic distance in mammals and fruit flies (Zou and Zhang 2015). Our analysis spans larger phylogenetic distances than that of Goldstein et al. (2015); still, most of the observed effect is local (fig. 4). On the basis of the data from different sources, and assuming a two-state fitness space such that each amino acid variant at a particular amino acid site can be either “prohibited” or “permitted”, the rate at which a particular variant switches between these two states has been estimated as ~5 such switches per unit time required for a single amino acid substitution to occur at this site (Usmanova et al. 2015). In our data, the rate of SPFL change appears to vary widely between proteins, as the time necessary for the P/D ratio to reach 1 varies between 0.6 for ATP6 and 2.1 for ND6. It also is strongly dependent on the size and the shape of the phylogeny. Still, in our data, the rate of SPFL changes has roughly the same scale as the rate of amino acid evolution (fig. 4).

In summary, our results provide an unambiguous demonstration of the change in the fitness landscapes of amino acid sites of mitochondrial proteins with time, supporting previous conjectures that such landscapes are dynamic in this dataset (Breen et al. 2012; Breen et al. 2013). Whether these changes are driven by changes in the intra-protein or inter-protein genomic context between species, or by environmental changes, remains a subject for future research.

Acknowledgements

This work was supported by the Russian Science Foundation grant no.14-50-00150.

We thank Shamil Sunyaev, Alexey Kondrashov, Alexander Favorov, Andrey Mironov, Dmitry Pervouchine, Vladimir Seplyarskiy, Sergey Naumenko, Elena Nabieva, Ivan Cytovich, David McCandlish, Joshua Plotkin, Richard Goldstein and Dinara Usmanova for valuable comments.

References

Bazykin GA. 2015. Changing preferences: deformation of single position amino acid fitness landscapes and evolution of proteins. Biol. Lett. 11.

Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. 2012. Epistasis as the primary factor in molecular evolution. Nature 490:535–538.

Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. 2013. Reply to McCandlish et al. Nature 497:E1–E2; discussion E2–E3.

Fitch WM. 1971. Rate of change of concomitantly variable codons. J. Mol. Evol. 1:84–96.

Fitch WM, Markowitz E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4:579–593.

Goldstein RA, Pollard ST, Shah SD, Pollock DD. 2015. Nonadaptive Amino Acid Convergence Rates Decrease over Time. Mol. Biol. Evol. 32:1373–1381.

Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinforma. Oxf. Engl. 23:127–128.

Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol. Biol. Evol. 19:1–7.

McCandlish DM, Rajon E, Shah P, Ding Y, Plotkin JB. 2013. The role of epistasis in protein evolution. Nature 497:E1–E2; discussion E2–E3.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 8:e1002764.

Naumenko SA, Kondrashov AS, Bazykin GA. 2012. Fitness conferred by replaced amino acids declines with time. Biol. Lett. 8:825–828.

Povolotskaya IS, Kondrashov FA. 2010. Sequence space and the ongoing expansion of the protein universe. Nature 465:922–926.

Rodrigue N. 2013. On the statistical interpretation of site-specific variables in phylogeny-based substitution models. Genetics 193:557–564.

Rogozin IB, Thomson K, Csürös M, Carmel L, Koonin EV. 2008. Homoplasy in genome-wide analysis of rare amino acid replacements: the molecular-evolutionary basis for Vavilov’s law of homologous series. Biol. Direct 3:7.

Roure B, Philippe H. 2011. Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. BMC Evol. Biol. 11:17.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinforma. Oxf. Engl. 30:1312–1313.

Tamuri AU, Dos Reis M, Hay AJ, Goldstein RA. 2009. Identifying changes in selective constraints: host shifts in influenza. PLoS Comput. Biol. 5:e1000564.

Usmanova DR, Ferretti L, Povolotskaya IS, Vlasov PK, Kondrashov FA. 2015. A model of substitution trajectories in sequence space and long-term protein evolution. Mol. Biol. Evol. 32:542–554.

Wiegand T, A. Moloney K. 2004. Rings, circles, and null-models for point pattern analysis in ecology. Oikos 104:209–229.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. CABIOS 13:555–556.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19:908–917.

Zou Z, Zhang J. 2015. Are Convergent and Parallel Amino Acid Substitutions in Protein Evolution More Prevalent Than Neutral Expectations? Mol. Biol. Evol. 32:2085–2096.

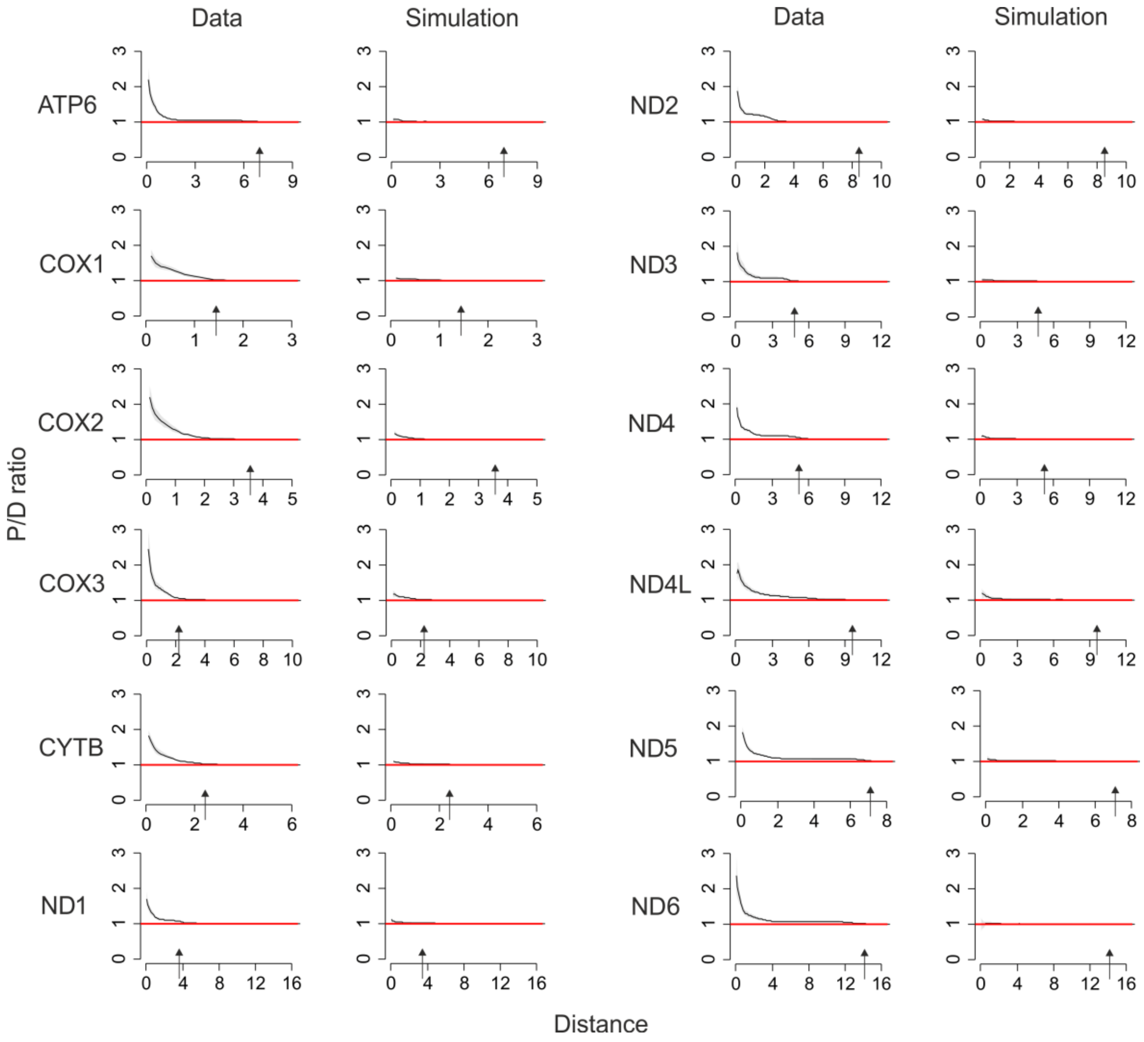
Supplementary materials

Supplementary Table 1.

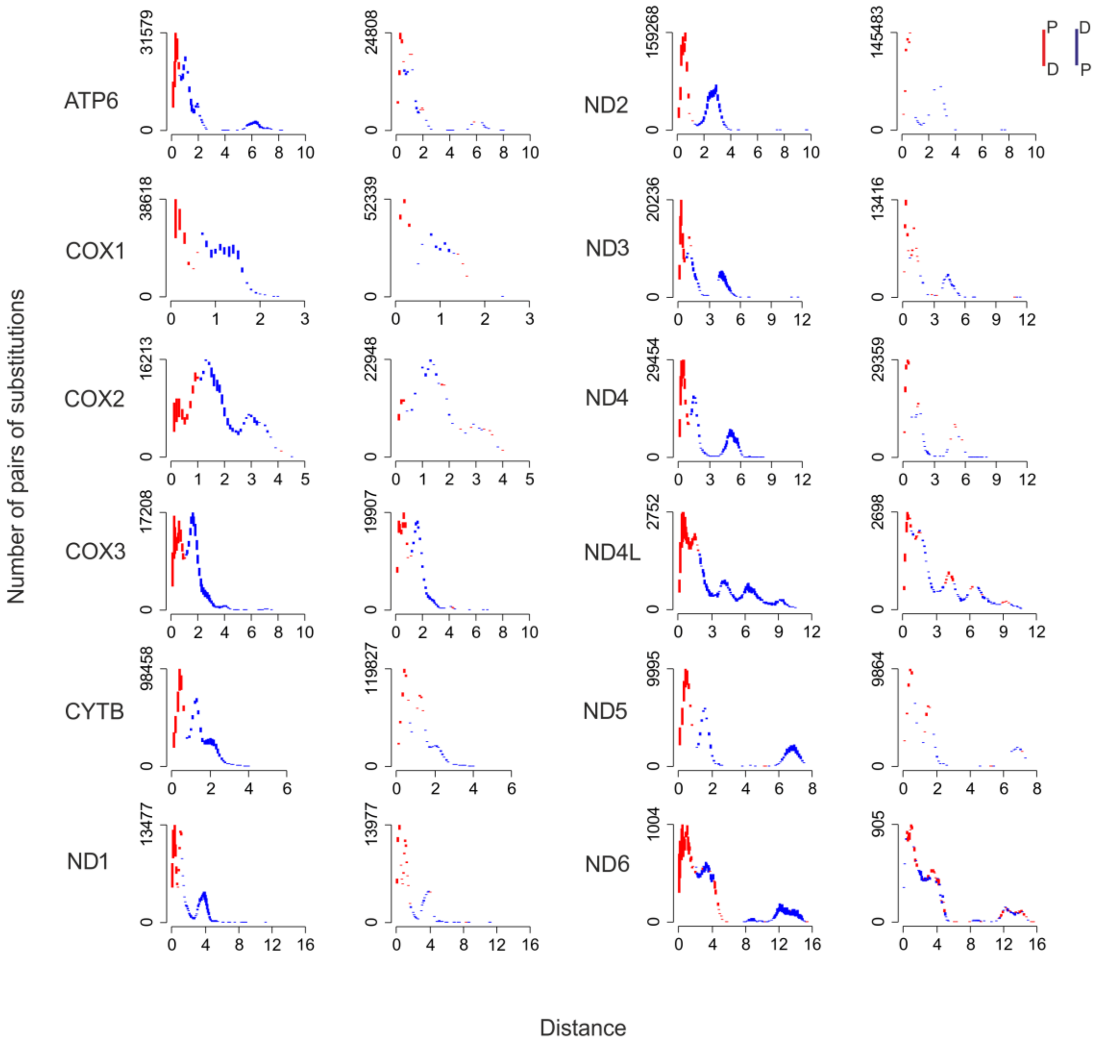
Analysis of homoplasies in metazoan phylogenies.

Gene	**Sites**	**Homoplasy informative sites***
ATP6	186	172
COX1	404	317
COX2	165	154
COX3	198	173
CYTB	327	321
ND1	253	227
ND2	299	288
ND3	94	87
ND4	392	348
ND4L	82	79
ND5	516	404
ND6	119	111

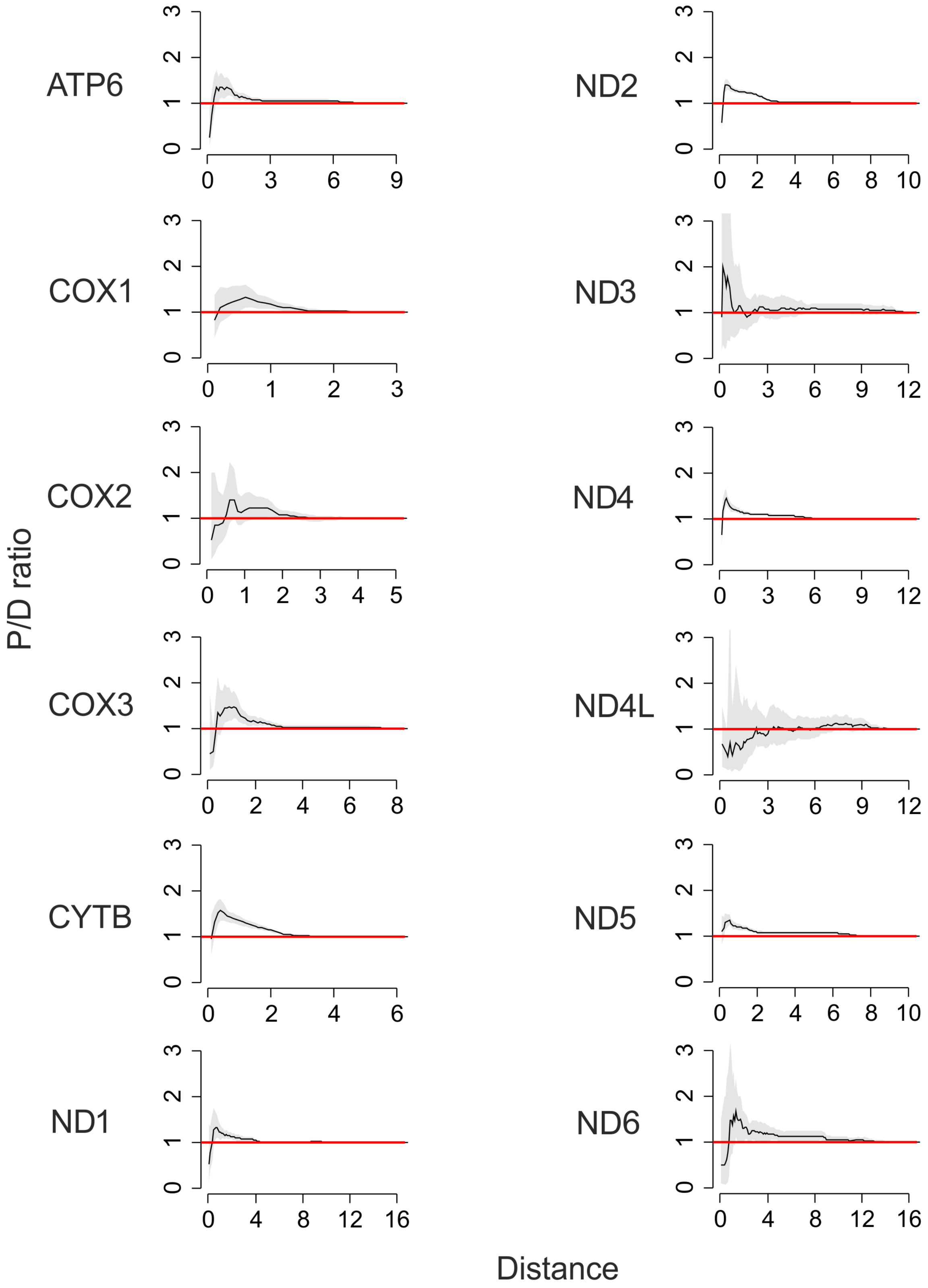
* Homoplasy informative sites are those that carry at least one pair of parallel substitutions, and at least one pair of divergent substitutions, of the same amino acid.



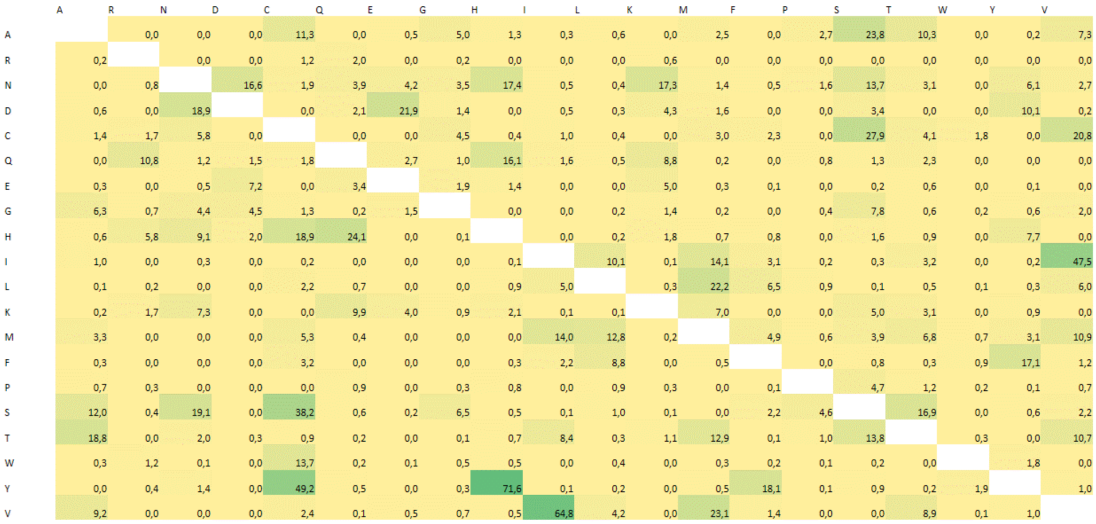
Supplementary figure 1. Higher fraction of parallel substitutions between closely related species in metazoan phylogenies. Horizontal axis, distance between branches carrying the substitutions, measured in numbers of amino acid substitutions per site (0.1 distance windows). Vertical axis, P/D ratios for substitutions at distances not exceeding this distance (unlike Figure 1, where the ratios for the substitutions falling into this distance bin are shown). Black line, mean; grey confidence band, 95% confidence interval obtained from 1000 bootstrapping trials. The red line shows the expected P/D ratio of 1. Arrows represent the distance between human and *Drosophila*.



Supplementary figure 2. Numbers of pairs of parallel (P) and divergent (D) substitutions. For each distance window of size 0.1, the two ends of the bar correspond to the average (across 1000 bootstrap replicates) numbers of P and D, so that bar length is equal to the absolute value of their difference. Red bars correspond to P>D (so that the top of the bar corresponds to P, and the bottom of the bar, to D), and blue bars correspond to P<D (so that the top of the bar corresponds to D, and the bottom of the bar, to P).



Supplementary figure 3. Higher fraction of parallel substitutions between closely related species in metazoan phylogenies, for robust phylogenetic branches. The figure is similar to S1, but only substitutions on branches with bootstrap support >90 were considered.



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A		0,0	0,0	0,0	11,3	0,0	0,5	5,0	1,3	0,3	0,6	0,0	2,5	0,0	2,7	23,8	10,3	0,0	0,2	7,3
R	0,2		0,0	0,0	1,2	2,0	0,0	0,2	0,0	0,0	0,0	0,6	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
N	0,0	0,8		16,6	1,9	3,9	4,2	3,5	17,4	0,5	0,4	17,3	1,4	0,5	1,6	13,7	3,1	0,0	6,1	2,7
D	0,6	0,0	18,9		0,0	2,1	21,9	1,4	0,0	0,5	0,3	4,3	1,6	0,0	0,0	3,4	0,0	0,0	10,1	0,2
C	1,4	1,7	5,8	0,0		0,0	0,0	4,5	0,4	1,0	0,4	0,0	3,0	2,3	0,0	27,9	4,1	1,8	0,0	20,8
Q	0,0	10,8	1,2	1,5	1,8		2,7	1,0	16,1	1,6	0,5	8,8	0,2	0,0	0,8	1,3	2,3	0,0	0,0	0,0
E	0,3	0,0	0,5	7,2	0,0	3,4		1,9	1,4	0,0	0,0	5,0	0,3	0,1	0,0	0,2	0,6	0,0	0,1	0,0
G	6,3	0,7	4,4	4,5	1,3	0,2	1,5		0,0	0,0	0,2	1,4	0,2	0,0	0,4	7,8	0,6	0,2	0,6	2,0
H	0,6	5,8	9,1	2,0	18,9	24,1	0,0	0,1		0,0	0,2	1,8	0,7	0,8	0,0	1,6	0,9	0,0	7,7	0,0
I	1,0	0,0	0,3	0,0	0,2	0,0	0,0	0,0	0,1		10,1	0,1	14,1	3,1	0,2	0,3	3,2	0,0	0,2	47,5
L	0,1	0,2	0,0	0,0	2,2	0,7	0,0	0,0	0,9	5,0		0,3	22,2	6,5	0,9	0,1	0,5	0,1	0,3	6,0
K	0,2	1,7	7,3	0,0	0,0	9,9	4,0	0,9	2,1	0,1	0,1		7,0	0,0	0,0	5,0	3,1	0,0	0,9	0,0
M	3,3	0,0	0,0	0,0	5,3	0,4	0,0	0,0	0,0	14,0	12,8	0,2		4,9	0,6	3,9	6,8	0,7	3,1	10,9
F	0,3	0,0	0,0	0,0	3,2	0,0	0,0	0,0	0,3	2,2	8,8	0,0	0,5		0,0	0,8	0,3	0,9	17,1	1,2
P	0,7	0,3	0,0	0,0	0,0	0,9	0,0	0,3	0,8	0,0	0,9	0,3	0,0	0,1		4,7	1,2	0,2	0,1	0,7
S	12,0	0,4	19,1	0,0	38,2	0,6	0,2	6,5	0,5	0,1	1,0	0,1	0,0	2,2	4,6		16,9	0,0	0,6	2,2
T	18,8	0,0	2,0	0,3	0,9	0,2	0,0	0,1	0,7	8,4	0,3	1,1	12,9	0,1	1,0	13,8		0,3	0,0	10,7
W	0,3	1,2	0,1	0,0	13,7	0,2	0,1	0,5	0,5	0,0	0,4	0,0	0,3	0,2	0,1	0,2	0,0		1,8	0,0
Y	0,0	0,4	1,4	0,0	49,2	0,5	0,0	0,3	71,6	0,1	0,2	0,0	0,5	18,1	0,1	0,9	0,2	1,9		1,0
V	9,2	0,0	0,0	0,0	2,4	0,1	0,5	0,7	0,5	64,8	4,2	0,0	23,1	1,4	0,0	0,0	8,9	0,1	1,0	

Supplementary figure 4. Substitution matrices for a combined phylogenetic tree constructed from ATP6 and ND1 genes for vertebrate (below the diagonal) and invertebrate (above the diagonal) subtrees. The color shade corresponds to the relative frequency of the mutation.