

1 **Parallel evolution of metazoan mitochondrial proteins**

2 **Galya V. Klink<sup>1</sup> and Georgii A. Bazykin<sup>1,2,3,\*</sup>**

3 <sup>1</sup>Institute for Information Transmission Problems (Kharkevich Institute) of  
4 the Russian Academy of Sciences, Moscow 127051, Russia

5 <sup>2</sup>Skolkovo Institute of Science and Technology, Skolkovo, 143025, Russia

6 <sup>3</sup>Lomonosov Moscow State University, Moscow 119234, Russia

7

8 \*Corresponding author: E-mail: [gbazykin@iitp.ru](mailto:gbazykin@iitp.ru)

9

10 **Abstract**

11 Amino acid propensities at amino acid sites change with time due to  
12 epistatic interactions or changing environment, affecting the probabilities of  
13 fixation of different amino acids. Such changes should lead to an increased rate  
14 of homoplasies (reversals, parallelisms, and convergences) at closely related  
15 species. Here, we reconstruct the phylogeny of twelve mitochondrial proteins  
16 from several thousand metazoan species, and measure the phylogenetic  
17 distances between branches at which either the same allele originated repeatedly  
18 due to homoplasies, or different alleles originated due to divergent substitutions.  
19 The mean phylogenetic distance between parallel substitutions is ~20% lower  
20 than the mean phylogenetic distance between divergent substitutions, indicating  
21 that a variant fixed in a species is more likely to be deleterious in a more

1 phylogenetically remote species, compared to a more closely related species.  
2 These findings are robust to artefacts of phylogenetic reconstruction or of  
3 pooling of sites from different conservation classes or functional groups, and  
4 imply that single-position fitness landscapes change at rates similar to rates of  
5 amino acid changes.

6 **Key words:** fitness landscape, epistasis, parallel substitutions,  
7 heteropecilly, mitochondria

8

## 9 **Introduction**

10 Amino acid preferences at a site, or single-position fitness landscape  
11 (SPFL, Bazykin 2015), change in the course of evolution, so that a variant  
12 conferring high fitness in one species may confer low fitness in another, either  
13 due to changes at interacting genomic sites or in the environment. These  
14 changes can be observed through phylogenetic patterns, in particular, through a  
15 non-uniform distribution of amino acid substitutions giving rise to a particular  
16 variant (homoplasies) along the phylogeny. Indeed, when a certain amino acid  
17 repeatedly arises at a particular site in a certain phylogenetic clade, but is never  
18 observed at this site in another clade, this implies that the relative fitness  
19 conferred by this variant in the former clade is higher. Different types of  
20 homoplasies – reversals, parallelisms and convergencies – have been found to be  
21 clustered on the phylogenies (Rogozin et al. 2008; Povolotskaya and

1 Kondrashov 2010; Naumenko et al. 2012; Goldstein et al. 2015; Zou and Zhang  
2 2015), and an attempt has been made to estimate the rate at which SPFLs change  
3 from such data (Usmanova et al. 2015).

4 SPFL changes in metazoan mitochondrial proteins were previously inferred  
5 from amino acid usage patterns (Breen et al. 2012), but this approach has been  
6 criticized as sensitive to the underlying assumptions regarding fitness  
7 distributions (McCandlish et al. 2013). Here, we develop an approach for the  
8 study of phylogenetic clustering of homoplasies at individual protein sites, and  
9 apply it to deep alignments of mitochondrial proteins of metazoans (Breen et al.  
10 2012) together with their reconstructed phylogenies. Our approach compares the  
11 distributions of distances between parallel and divergent substitutions to infer  
12 robustly changes in relative fitness of different variants at a site between  
13 branches of the phylogenetic tree.

14

## 15 **Materials and Methods**

### 16 **Alignment and phylogeny**

17 We obtained multiple-species alignments of 12 mitochondrial proteins of  
18 metazoans from (Breen et al. 2012), and analyzed alignment columns with fewer  
19 than 1% gaps (which comprised 77% of all sites). As there is no accepted  
20 phylogenetic tree for this large and diverse set of species spanning a wide range  
21 of phylogenetic distances, we took a hybrid approach to reconstructing their

1 phylogeny. First, we constrained the tree topology using the curated taxonomy-  
2 based phylogeny of the ITOL (Interactive tree of life) project (Letunic and Bork  
3 2007). By requiring the presence of each species in the ITOL database, we were  
4 left with >900 metazoan species for each protein (Table 1). The resulting  
5 topology was not fully resolved, and contained multifurcations. We then used  
6 RAxML 8.0.0 (Stamatakis 2014) under the GTR- $\Gamma$  model to resolve  
7 multifurcations and to estimate the branch lengths. Finally, ancestral states were  
8 reconstructed using codeml program of the PAML package (Yang 1997) under  
9 the substitution matrix and the value of the parameter alpha of the gamma  
10 distribution inferred by RAxML.

11 Independently, using the same methods and parameters, we reconstructed a  
12 joint phylogeny of 3586 chordate, 586 non-chordate and 178 fungal species (for  
13 a total of 4350 species) based on amino acid sequences of five concatenated  
14 mitochondrial genes, covering a total of 1524 amino acid positions.

15 Transmembrane and non-membrane sites were obtained from UniProt  
16 database (<http://www.uniprot.org/>).

### 17 **Clustering of substitutions on a phylogeny**

18 Using the inferred states of amino acid sites at each node, we inferred the  
19 positions of all substitutions at all protein sites on the phylogenies of the  
20 corresponding proteins. For each ancestral amino acid at a site, we defined  
21 parallel substitutions as those giving rise to the same derived amino acid, and  
22 divergent substitutions, as those giving rise to different derived amino acids

1 (fig. 1A). We considered only those pairs of substitutions that happened in  
2 phylogenetically independent branches, i.e., such that one was not ancestral to  
3 the other. For subsequent analyses, we used only homoplasy-informative sites,  
4 i.e., sites that have at least one pair of parallel substitutions and one pair of  
5 divergent substitutions from the same ancestral amino acid. The phylogenetic  
6 distance between a pair of substitutions was defined as the distance (measured in  
7 the number of amino acid substitutions per amino acid site inferred by RAxML)  
8 between the centers of the edges where those substitutions have occurred, i.e.,  
9 the sum of the distances from the centers of these edges to the last common  
10 ancestor of the two substitutions (fig. 1A). Alternatively, as a proxy for the site-  
11 specific evolutionary distance, we multiplied the phylogenetic distance by the  
12 codeml estimate of the site-specific substitution rate.

13 To compare the distances between parallel and divergent substitutions  
14 while circumventing the potential biases associated with pooling sites and amino  
15 acids with different properties (see below), we subsampled the pairs of parallel  
16 and divergent substitutions, and analyzed the distances in this subset. For this,  
17 for each ancestral amino acid at homoplasy-informative sites, we picked  
18 randomly  $\min(N_{\text{paral}}, N_{\text{diverg}})$  pairs of parallel substitutions and the same number  
19 of pairs of divergent substitutions, where  $N_{\text{paral}}$  and  $N_{\text{diverg}}$  are the numbers of  
20 parallel and divergent substitutions originating from this amino acid. We  
21 repeated this procedure for all ancestral amino acids at all sites, thus obtaining  
22 two equal-sized subsamples of parallel and divergent substitutions, and

1 measured all distances in these resulting subsamples. The parallel to divergent  
2 ratio (P/D) for each distance bin was calculated by dividing the number of  
3 parallel pairs of substitutions by the number of divergent pairs of substitutions  
4 that had occurred at a distance from each other falling into this bin. This statistic  
5 is closely related to the O-ring statistic widely used in spatial ecology to  
6 measure aggregation in communities (Wiegand and A. Moloney 2004).

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

10

### 11 **Robustness of tree shape and ancestral states reconstruction**

12 We identified a set of high-confidence pairs of substitutions, defined as  
13 follows. For each branch of the phylogenetic tree, we obtained the bootstrap  
14 support value in 100 bootstrap replicates using RAxML. A pair of substitutions  
15 was considered high-confidence when (i) at least one node between substitutions  
16 had 100% bootstrap support, ensuring the robustness of these nodes; and (ii) for  
17 each substitution from the pair, the maximum likelihood estimate for amino  
18 acids in ancestral and derived nodes was equal to 1, ensuring the robustness of  
19 ancestral state reconstruction.

20

### 21 **Simulated evolution**

1 For each gene, we simulated amino acid evolution using the evolver  
2 program of the PAML package (Yang 1997), under the empirical\_F model and  
3 discrete-gamma distributed rates between sites. The phylogenetic tree,  
4 substitution matrix, alpha parameter and number of categories for discrete  
5 gamma of the gamma-distribution were obtained from the output of RAxML for  
6 the corresponding gene. From the amino acids thus simulated for the leaves of  
7 the tree (i.e., extant species), we then reconstructed the ancestral states using  
8 codeml under the same parameters.

9

## 10 **Simulated evolution under different substitution matrices**

11 To test the effect of differences in substitution matrices between clades due  
12 to clade-specific biases, we obtained individual RAxML-generated matrices for  
13 each of the three major groups of species in the joint 5-gene phylogeny:  
14 chordates, non-chordates and fungi, using the same methods as above. We then  
15 used these matrices to simulate evolution of the corresponding groups of species  
16 of the joint 4350-species tree, and used generated sequences for the analysis.

17

## 18 **Results**

### 19 **Phylogenetic clustering as evidence for SPFL changes**

20 We devised an approach for analysis of the clustering of parallel  
21 substitutions at a site. While conceptually related to the previous methods

1 (Povolotskaya and Kondrashov 2010; Goldstein et al. 2015; Zou and Zhang  
2 2015), it is designed to be robust to other specifics of the phylogenetic  
3 distribution of substitutions, and to control for any potential biases that can arise  
4 from pooling sites with different properties. Since it is difficult to obtain robust  
5 evidence for SPFL changes for an individual amino acid site even using large  
6 numbers of species, getting a significant signal of SPFL changes requires  
7 pooling different amino acid sites. The problem is that these sites may differ in  
8 their properties, and such differences may lead to artefactual evidence for SPFL  
9 changes, for the following reasons.

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

22 *Figure 1*



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

17       We circumvent this problem by resampling matched sets of parallel and  
18 divergent substitutions. Specifically, at each homoplasy-informative site (see  
19 Methods), we consider different ancestral amino acids separately. For each  
20 ancestral amino acid B, we subsample our sets of pairs of substitutions: for each  
21 pair of parallel ( $B \rightarrow A$ ,  $B \rightarrow A$ ) substitutions, we randomly pick exactly one pair  
22 of divergent ( $B \rightarrow A$ ,  $B \rightarrow X$ ) substitutions from the same site. Finally, we pool

1 together these subsamples from different sites, and analyze distances between  
2 parallel and divergent substitutions in this pooled set. This approach controls for  
3 any possible biases associated with differences in phylogenetic distributions of  
4 different ancestral amino acids. From the resulting subsets of distances, we  
5 calculate the ratio of the numbers of parallel to divergent substitutions (P/D) for  
6 each distance bin (see Methods).

7

## 8 **Parallel substitutions in mitochondrial proteins are phylogenetically** 9 **clustered**

10 We applied this approach to the phylogenetic trees of 12 orthologous  
11 mitochondrial proteins of metazoans, each including >900 species. At the vast  
12 majority of sites, we observe many amino acid variants, in line with (Breen et al.  
13 2012). By reconstructing the ancestral states and substitutions at each site, we  
14 observe that most of these variants have originated more than once, allowing us  
15 to study the phylogenetic distribution of homoplasies in detail (Table 1 and  
16 supplementary table 1, Supplementary Material online).

### 17 *Table 1*

18

19 We observe an excess of parallel substitutions for species at small  
20 phylogenetic distances from each other (fig. 2-3), in line with the previous  
21 findings in vertebrates that used a smaller dataset (Goldstein et al. 2015). The  
22 P/D ratio is ~1.7 to 2.5 at phylogenetic distances less than 0.1, but drops to ~1

1 rapidly for larger distances (fig. 4 and supplementary fig. 1, Supplementary  
2 Material online). In simulated data, only a very weak decrease in the P/D ratio  
3 was observed, which is possibly attributable to minor biases in phylogenetic  
4 reconstruction.

5 *Figure 2*

6 *Figure 3*

7 *Figure 4*

8 We also measured P/D ratios for phylogenetic distances normalized by site-  
9 specific evolutionary rates (see Methods) and obtained similar plots  
10 (Supplementary fig. 2, Supplementary Material online). We also asked whether  
11 the mean P/D distance is different between sites with different evolutionary  
12 rates, but saw no systematic differences (Supplementary fig. 3, Supplementary  
13 Material online). These findings suggest that the P/D ratios are more sensitive to  
14 the evolutionary distance spanned by the species rather than by the individual  
15 site. Finally, we observed similar effect in trans-membrane and in non-  
16 membrane residues. In most proteins, the effect is slightly weaker in trans-  
17 membrane residues, but this difference is extremely weak (Supplementary fig. 4,  
18 Supplementary Material online).

19 The rate at which the P/D ratio declines with phylogenetic distance varies a  
20 lot between genes. We asked whether this difference has to do with the intrinsic  
21 rate of protein evolution, which also varied strongly between genes. We used the  
22 number of substitutions between human and *Drosophila* as the proxy for the

1 evolutionary rate of the protein, with higher values corresponding to rapidly  
2 evolving genes; and the phylogenetic distance at which the P/D ratio (which is  
3 initially always larger than one) reaches one, as the proxy for the rate of the  
4 decline of the P/D ratio, with higher values corresponding to a slower decline.  
5 Among the 10 analyzed genes for which sequences both for human and  
6 *Drosophila* were available, the P/D decline appeared to be somewhat faster in  
7 fast-evolving genes, although this trend was not significant (Spearman's test:  
8  $R=0.53$ ,  $p=0.11$ ).

9 *Figure 5*

10

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

13 Conceivably, the decline of the P/D ratio could be an artefact of erroneous  
14 phylogenetic reconstruction. Indeed, if a clade is erroneously split on a  
15 phylogeny, synapomorphies (shared derived character states) may be mistaken  
16 for parallel substitutions, and this is more likely for closely related species  
17 (Mendes et al. 2016). We tested the contribution of such artefacts by performing  
18 our analyses only on high-confidence pairs of substitutions (see Methods). Our  
19 definition of this set was conservative, because branches with parallel  
20 substitutions are expected to have a reduced bootstrap support, as such  
21 substitutions cause attraction of the branches where they occur in phylogenetic  
22 reconstruction. Indeed, in the Breen et al. dataset, this procedure discarded the

1 vast majority of parallel substitutions at very small phylogenetic distances,  
2 leaving us with too little data. To circumvent this, we reconstructed a 5-gene,  
3 4350-species phylogeny (see Materials and Methods for details). Similarly to the  
4 main analysis, the P/D ratio declined monotonically with distance between  
5 substitutions in this dataset (Figure 6), implying that the excess of homoplasies  
6 at short phylogenetic distances is unlikely to be an artefact of phylogenetic  
7 reconstruction.

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

13 Can the changes in substitution rates with phylogenetic distance be  
14 explained by changes in mutation rates? There can be two scenarios for this.  
15 First, the mitochondrial mutational spectra could differ between clades,  
16 potentially leading to differences in the rate at which a particular mutation  
17 occurs. If the mutation rate corresponding to a particular substitution is much  
18 higher in a particular clade, compared to the rest of the phylogeny, this may lead  
19 to an excess of homoplasies falling into this clade. To ask whether this  
20 mechanism contributes, we simulated evolution of the three major clades of the  
21 4350-species phylogeny using an independent substitution matrix for each clade  
22 (see Materials and Methods for details) and performed our analysis for the

1 whole phylogeny using simulated sequences. Results for this simulation are  
2 indistinguishable from the simulation constructed with one matrix for all clades  
3 (Fig. 6), implying that this mechanism is unlikely to cause the observed pattern.  
4 Moreover, most of the change in the P/D ratio occurs at very small phylogenetic  
5 distances (fig. 4), where the mutation matrices are very similar, and unlikely to  
6 contribute to our effect.

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

17       Finally, we analyzed the distribution of the pairs across the phylogenetic  
18 tree. Both in data and in simulations, substitutions with high (or low) P/D are  
19 clustered on the tree. The distribution of parallel (as well as divergent)  
20 substitutions over the tree is non-uniform mainly because of differences in  
21 branch lengths: longer branches carry more substitutions of all kinds. To test if  
22 there is an excess of branches with more than expected parallel substitutions, we

1 plotted the distribution of pairs of branches by the number of parallel  
2 substitutions that had occurred in them (Supplementary figure 5). We see no  
3 systematic differences between the distributions obtained from the data and from  
4 simulations, implying that our results are not driven by clustering of  
5 substitutions in some of the branch pairs. Moreover, in branch pairs that carried  
6 many substitution pairs in them, both branches tended to be long, leading to  
7 large distances between the two substitutions in a pair; while most of the effect  
8 is observed at small distances (Figure 4).

9 For three of the genes, we also randomly picked and examined visually 20  
10 pairs of parallel and 20 pairs of divergent substitutions among those with  
11 distances  $<0.1$  between them. Parallel and divergent pairs in data and in  
12 simulations were clustered in the same parts of the tree, demonstrating that this  
13 effect is likely to be determined by the shape of the tree. Specifically, they were  
14 clustered in the regions of the tree with many short branches, which are the only  
15 ones contributing to the phylogenetically close pairs of substitutions in Figure 4.

16

## 17 **Discussion**

18 The rate at which a specific substitution occurs is a monotonic function of  
19 fitness differences between the descendant and the ancestral variants, and  
20 changes in the substitution rates in the course of evolution indicate that these  
21 fitness differences, and thus the SPFL, change. Obtaining the entire substitution

1 matrix for individual amino acid sites is problematic even given hundreds of  
2 species (Rodrigue 2013); still, the changes in the relative substitution rates can  
3 be inferred using some summary statistics. The change in the extent of  
4 parallelism in the course of evolution is one such convenient statistic: as a  
5 substitution becomes more deleterious, its rate decreases, and it becomes more  
6 sparsely distributed on the phylogeny.

7 We observe that parallel substitutions in the evolution of metazoan  
8 mitochondrial proteins are phylogenetically clustered; i.e., that such  
9 substitutions are more likely to occur in the phylogenetic vicinity of each other,  
10 compared with divergent substitutions. As a result, the distance between two  
11 parallel substitutions on the phylogenetic tree is, on average, ~20% lower than  
12 the distance between divergent substitutions, or than expected if their rate was  
13 constant across the tree (fig. 3). We show that these results are unlikely to be  
14 artefacts of phylogenetic reconstruction or of pooling together sites and amino  
15 acids with different properties. Our results cannot be explained by a simple  
16 covarion model, in which a site alternates between neutral and constrained  
17 (Fitch and Markowitz 1970; Fitch 1971), as the changes we observe are not  
18 associated with changes in the overall substitution rates. For the same reason,  
19 they also cannot be explained by a broader class of heterotachy models in which  
20 the overall rates of evolution of a site vary with time (Lopez et al. 2002; Yang  
21 and Nielsen 2002; Murrell et al. 2012), but require heteropecilly (Tamuri et al.  
22 2009; Roure and Philippe 2011), i.e., variation with time of rates of individual



1 substitutions. We show that these differences are unlikely to be caused by  
2 systematic gene- or genome-wide differences in substitution matrices between  
3 clades, which may result from differences in mutation patterns, selection for  
4 nucleotide or amino acid usage, or gene conversion.

5 Instead, they most likely reflect changes in single-position fitness  
6 landscapes (Mustonen and Lässig 2007; Bazykin 2015) that accumulate in the  
7 course of evolution. Indeed, site-specific differences in the rate of a substitution  
8 leading to a particular amino acid imply that the relative fitness of this amino  
9 acid relative to other amino acids at this site changes with time. Decrease in this  
10 frequency with phylogenetic distance may be caused by a decline in the fitness  
11 of this allele, and/or by an increase in the fitness of other alleles; it is hard to  
12 distinguish between these possibilities with the available data, although both  
13 factors likely play a role (Naumenko et al. 2012). Our single-site approach also  
14 prevents us from distinguishing between the possible causes of the SPFL  
15 changes: evolution of other sites (of the same or other proteins) involved in  
16 epistatic interactions with the focal site, environmental changes, or perhaps a  
17 combination of both.

18 The numbers of substitutions to the same or to another amino acid, i.e.  
19 convergent and divergent substitutions at different phylogenetic distances, have  
20 been used previously to characterize evolution. In ancient proteins, the rate of  
21 convergence monotonically decreases with phylogenetic distance, and half of  
22 the reversals were estimated to become forbidden after 10% protein divergence

1 (Povolotskaya and Kondrashov 2010). The ratio of the rates of convergent and  
2 divergent substitutions drops by more than twofold with phylogenetic distance  
3 within vertebrates (Goldstein et al. 2015). Similarly, the rate of convergence  
4 decreases with phylogenetic distance in mammals and fruit flies (Zou and Zhang  
5 2015). Our analysis spans larger phylogenetic distances than that of Goldstein et  
6 al. (2015); still, most of the observed effect is local (fig. 4). On the basis of the  
7 data from different sources, and assuming a two-state fitness space such that  
8 each amino acid variant at a particular amino acid site can be either “prohibited”  
9 or “permitted”, the rate at which a particular variant switches between these two  
10 states has been estimated as ~5 such switches per unit time required for a single  
11 amino acid substitution to occur at this site (Usmanova et al. 2015). In our data,  
12 the rate of SPFL change appears to vary widely between proteins, as the time  
13 necessary for the P/D ratio to reach 1 varies between 0.6 for ATP6 and 2.1 for  
14 ND6. It also is strongly dependent on the size and the shape of the phylogeny.  
15 Still, in our data, the rate of SPFL changes has roughly the same scale as the rate  
16 of amino acid evolution (fig. 4), which is consistent with the results of  
17 (Mustonen and Lässig 2007) who have shown that fluctuations in SPFLs of  
18 *Drosophila* proteins occur with rates comparable with neutral mutation rates.

19 In summary, our results allow to suggest that the fitness landscapes of  
20 amino acid sites of mitochondrial proteins change with time, supporting  
21 previous conjectures that such landscapes are dynamic in this dataset (Breen et  
22 al. 2012; Breen et al. 2013). Whether these changes are driven by changes in the

1 intra-protein or inter-protein genomic context between species, or by  
2 environmental changes, remains a subject for future research.

3

4

## 5 **Acknowledgements**

6 This work was supported by the Russian Science Foundation grant no.14-  
7 50-00150.

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

12

## 13 **References**

- 14 Bazykin GA. 2015. Changing preferences: deformation of single position amino  
15 acid fitness landscapes and evolution of proteins. *Biol. Lett.* 11.
- 16 Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. 2012. Epistasis  
17 as the primary factor in molecular evolution. *Nature* 490:535–538.
- 18 Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. 2013. Reply to  
19 McCandlish et al. *Nature* 497:E1–E2; discussion E2–E3.
- 20 Fitch WM. 1971. Rate of change of concomitantly variable codons. *J. Mol. Evol.*  
21 1:84–96.
- 22 Fitch WM, Markowitz E. 1970. An improved method for determining codon  
23 variability in a gene and its application to the rate of fixation of mutations in  
24 evolution. *Biochem. Genet.* 4:579–593.

- 1 Goldstein RA, Pollard ST, Shah SD, Pollock DD. 2015. Nonadaptive Amino Acid  
2 Convergence Rates Decrease over Time. *Mol. Biol. Evol.* 32:1373–1381.
- 3 Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for  
4 phylogenetic tree display and annotation. *Bioinforma. Oxf. Engl.* 23:127–  
5 128.
- 6 Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein  
7 evolution. *Mol. Biol. Evol.* 19:1–7.
- 8 McCandlish DM, Rajon E, Shah P, Ding Y, Plotkin JB. 2013. The role of epistasis  
9 in protein evolution. *Nature* 497:E1–E2; discussion E2–E3.
- 10 Mendes FK, Hahn Y, Hahn MW. 2016. Gene tree discordance can generate  
11 patterns of diminishing convergence over time. *Mol. Biol. Evol.*
- 12 Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL.  
13 2012. Detecting individual sites subject to episodic diversifying selection.  
14 *PLoS Genet.* 8:e1002764.
- 15 Mustonen V, Lässig M. 2007. Adaptations to fluctuating selection in *Drosophila*.  
16 *Proc. Natl. Acad. Sci. U. S. A.* 104:2277–2282.
- 17 Naumenko SA, Kondrashov AS, Bazykin GA. 2012. Fitness conferred by replaced  
18 amino acids declines with time. *Biol. Lett.* 8:825–828.
- 19 Povolotskaya IS, Kondrashov FA. 2010. Sequence space and the ongoing  
20 expansion of the protein universe. *Nature* 465:922–926.
- 21 Rodrigue N. 2013. On the statistical interpretation of site-specific variables in  
22 phylogeny-based substitution models. *Genetics* 193:557–564.
- 23 Rogozin IB, Thomson K, Csürös M, Carmel L, Koonin EV. 2008. Homoplasy in  
24 genome-wide analysis of rare amino acid replacements: the molecular-  
25 evolutionary basis for Vavilov’s law of homologous series. *Biol. Direct* 3:7.
- 26 Roure B, Philippe H. 2011. Site-specific time heterogeneity of the substitution  
27 process and its impact on phylogenetic inference. *BMC Evol. Biol.* 11:17.
- 28 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-  
29 analysis of large phylogenies. *Bioinforma. Oxf. Engl.* 30:1312–1313.
- 30 Tamuri AU, Dos Reis M, Hay AJ, Goldstein RA. 2009. Identifying changes in  
31 selective constraints: host shifts in influenza. *PLoS Comput. Biol.*  
32 5:e1000564.

- 1 Usmanova DR, Ferretti L, Povolotskaya IS, Vlasov PK, Kondrashov FA. 2015. A  
2 model of substitution trajectories in sequence space and long-term protein  
3 evolution. *Mol. Biol. Evol.* 32:542–554.
- 4 Wiegand T, A. Moloney K. 2004. Rings, circles, and null-models for point pattern  
5 analysis in ecology. *Oikos* 104:209–229.
- 6 Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum  
7 likelihood. *Comput. Appl. Biosci.* CABIOS 13:555–556.
- 8 Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular  
9 adaptation at individual sites along specific lineages. *Mol. Biol. Evol.*  
10 19:908–917.
- 11 Zou Z, Zhang J. 2015. Are Convergent and Parallel Amino Acid Substitutions in  
12 Protein Evolution More Prevalent Than Neutral Expectations? *Mol. Biol.*  
13 *Evol.* 32:2085–2096.

14

15 **Figure legends:**

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

1       **Figure 2.** Parallel and divergent substitutions at site 202 of ATP6 (NCBI  
2 reference sequence numbering for the human sequence). The ancestral variant  
3 (L) has experienced multiple substitutions, which are scattered throughout the  
4 phylogeny. However, the two parallel L→F substitutions occur in closely related  
5 species; the same is true for the three parallel L→V substitutions. Phylogenetic  
6 distances are in numbers of amino acid substitutions per site. The branches  
7 indicated with the blue waves are shortened by 1.2 distance units.

8       **Figure 3.** Ratios of phylogenetic distances between parallel and divergent  
9 substitutions in metazoan phylogenies. Values below 1 imply that the parallel  
10 substitutions are closer at the phylogeny to each other, compared to divergent  
11 substitutions. The bar height and the error bars represent respectively the median  
12 and the 95% confidence intervals obtained from 1,000 bootstrap replicates, and  
13 asterisks show the significance of difference from the one-to-one ratio (red line;  
14 \*\*\*,  $P < 0.001$ ; no asterisk,  $p > 0.05$ ). all, real data; accessible, real data only for  
15 substitutions from accessible amino acid pairs (see text); simulated, simulated  
16 data.

17       **Figure 4.** Higher fraction of parallel substitutions between closely related  
18 species. Horizontal axis, distance between branches carrying the substitutions,  
19 measured in numbers of amino acid substitutions per site (split into bins by  
20  $\log_2(\text{distance})$ ). Vertical axis, P/D ratios for substitutions at this distance. Black  
21 line, mean; grey confidence band, 95% confidence interval obtained from 1000

1 bootstrapping replicates. The red line shows the expected P/D ratio of 1. Arrows  
2 represent the distance between human and *Drosophila*.

3

4 **Figure 5.** Faster decline of the P/D ratio for rapidly evolving genes.  
5 Horizontal axis, gene-specific phylogenetic distance between *Homo sapiens* and  
6 *Drosophila simulans*. Vertical axis, phylogenetic distance at which the P/D ratio  
7 reaches 1. Each dot represents one gene, and the line represents the linear trend.  
8 Only the ten genes with available *Drosophila simulans* sequences were used.  
9 Error bars correspond to the 95% confidence interval for the distance at which  
10 the P/D ratio reaches 1, obtained by bootstrapping sites 1000 times.

11 **Figure 6.** Higher fraction of parallel substitutions between closely related  
12 species (A) and ratios of phylogenetic distances between parallel and divergent  
13 substitutions (B) in the 4350-species phylogeny, for high-confidence pairs of  
14 substitutions. sim3matr, simulation with independent substitution matrices for  
15 each clade (see text).

16 **Supplementary figure 1.** Numbers of pairs of parallel (P) and divergent  
17 (D) substitutions. For each distance window of size 0.1, the two ends of the bar  
18 correspond to the average (across 1000 bootstrap replicates) numbers of P and  
19 D, so that bar length is equal to the absolute value of their difference. Red bars  
20 correspond to  $P > D$  (so that the top of the bar corresponds to P, and the bottom of  
21 the bar, to D), and blue bars correspond to  $P < D$  (so that the top of the bar  
22 corresponds to D, and the bottom of the bar, to P).

1

2       **Supplementary figure 2.** Higher fraction of parallel substitutions between  
3 closely related species in metazoan phylogenies. Horizontal axis, distance  
4 between branches carrying the substitutions, measured in numbers of amino acid  
5 substitutions per site ( $\log_2$ distance bins). Vertical axis, P/D ratios for  
6 substitutions at distances falling into this distance bin. Black line, mean; grey  
7 confidence band, 95% confidence interval obtained from 1000 bootstrapping  
8 trials. The red line shows the expected P/D ratio of 1. The figure is similar to  
9 Figure 4, but phylogenetic distances between substitutions were measured using  
10 branch lengths normalized by site-specific evolutionary rates obtained by  
11 codeml.

12       **Supplementary figure 3.** Ratios of phylogenetic distances between  
13 parallel and divergent substitutions in metazoan phylogenies, for sites from  
14 different rate categories (1: slowest-evolving sites; 4: fastest-evolving sites).  
15 Notations are the same as in Figure 3.

16       **Supplementary figure 4.** Ratios of phylogenetic distances between  
17 parallel and divergent substitutions in metazoan phylogenies, for transmembrane  
18 (mem) and non-membrane (non-mem) sites. Red asterisks represent p-values of  
19 the 2-sided Wilcoxon rank sum test with the null-hypothesis of no difference  
20 between two types of sites (\*\*\*,  $P < 0.001$ ), with signs > or < indicating the  
21 direction of the difference. Other notations are as in Figure 3.



1        **Supplementary figure 5.** The distribution of pairs of branches by the  
2 fraction of all parallel substitutions that had occurred on them. Red, data; blue,  
3 simulation.

4

5

6

7

8

9

10

11

12

13

14

15

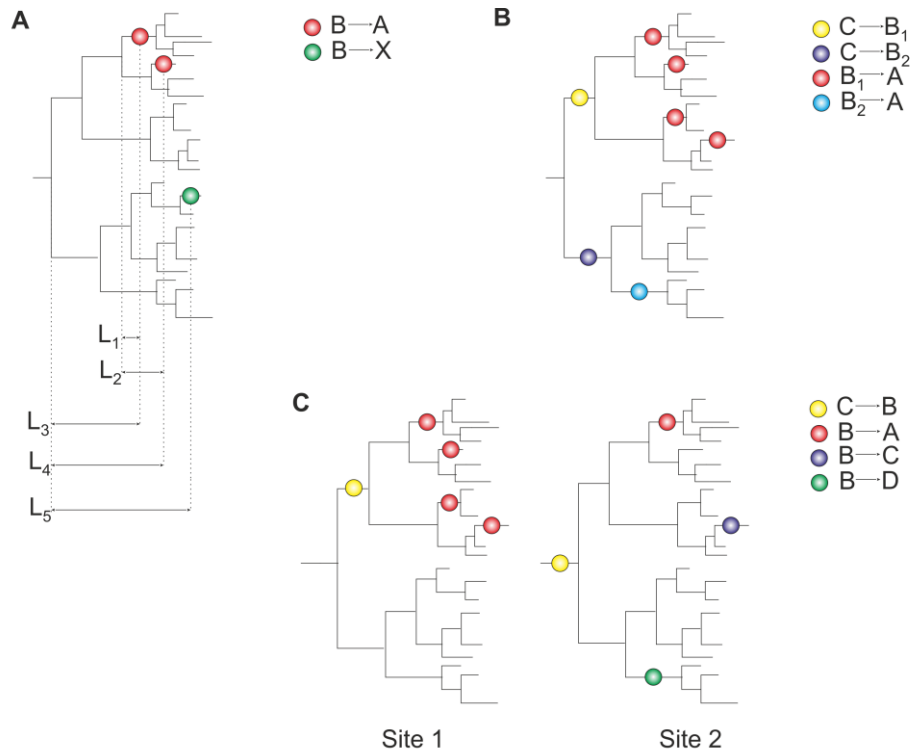
16

17    **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

1  
2  
3  
4  
5  
6  
7  
8  
9



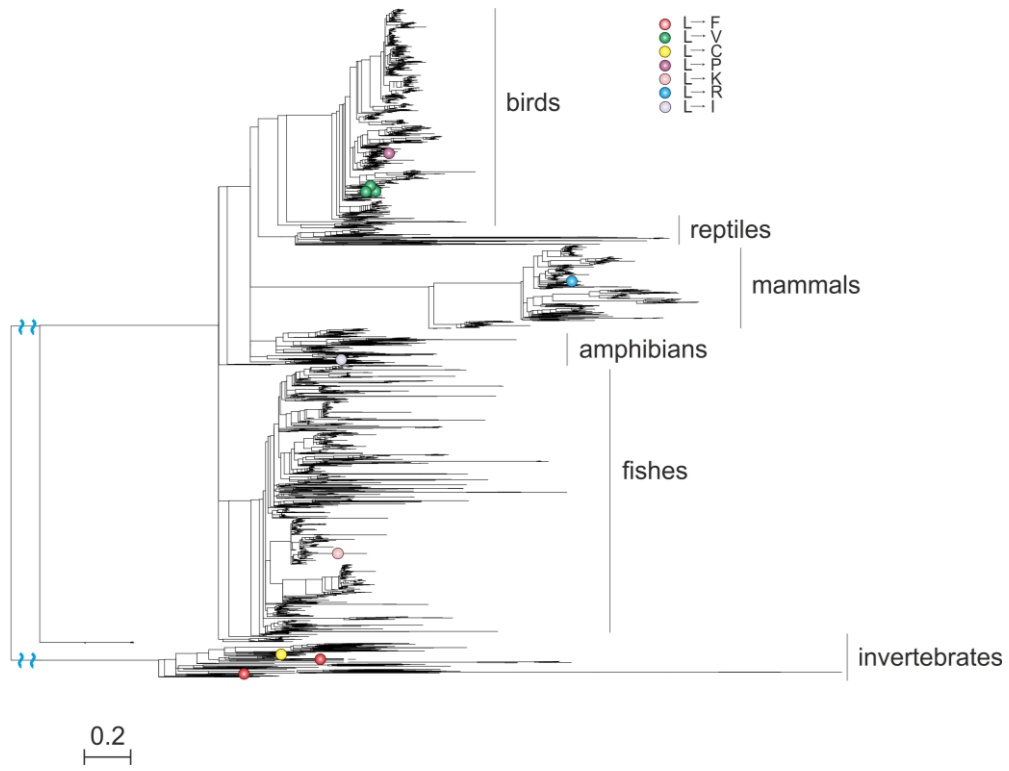
10  
11

1

2

3

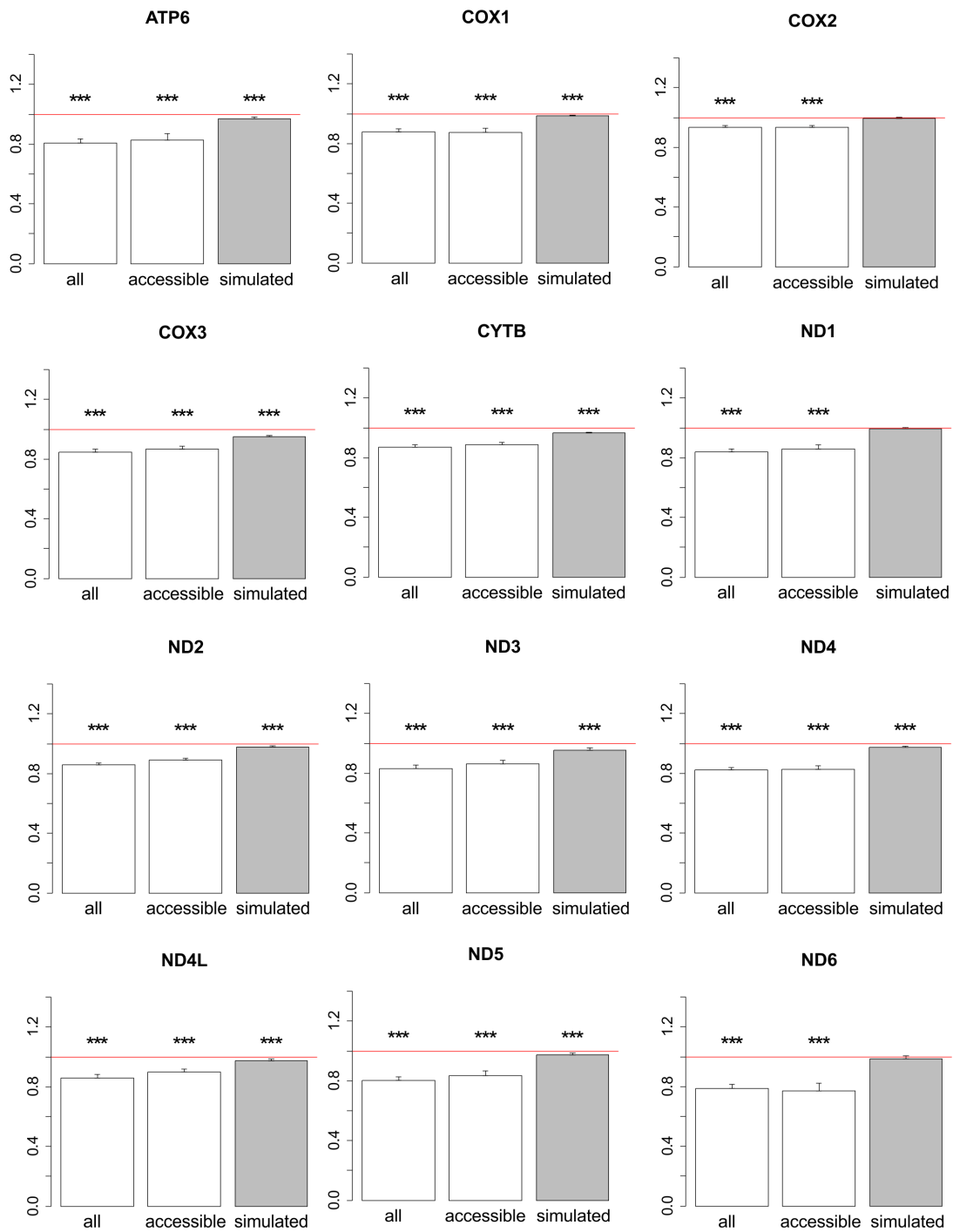
4



5

6

7

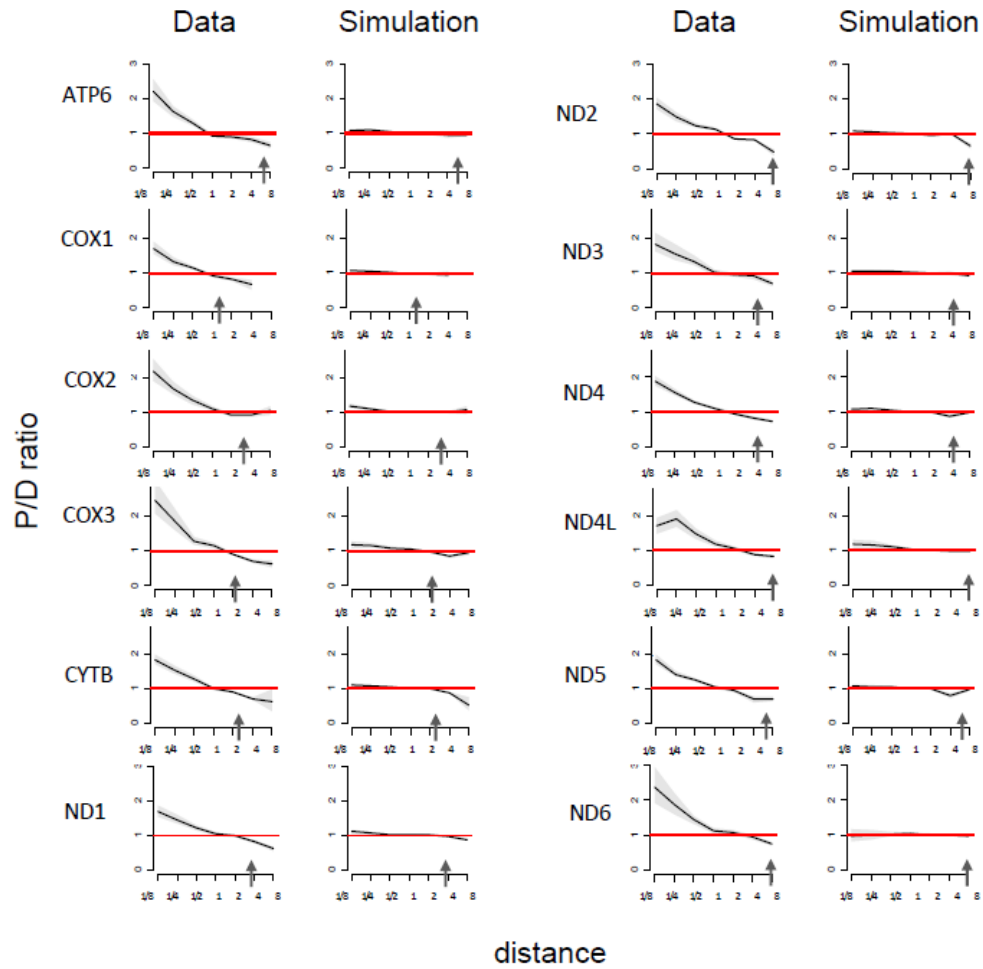


1

2

3

4



1

2

3

4

5

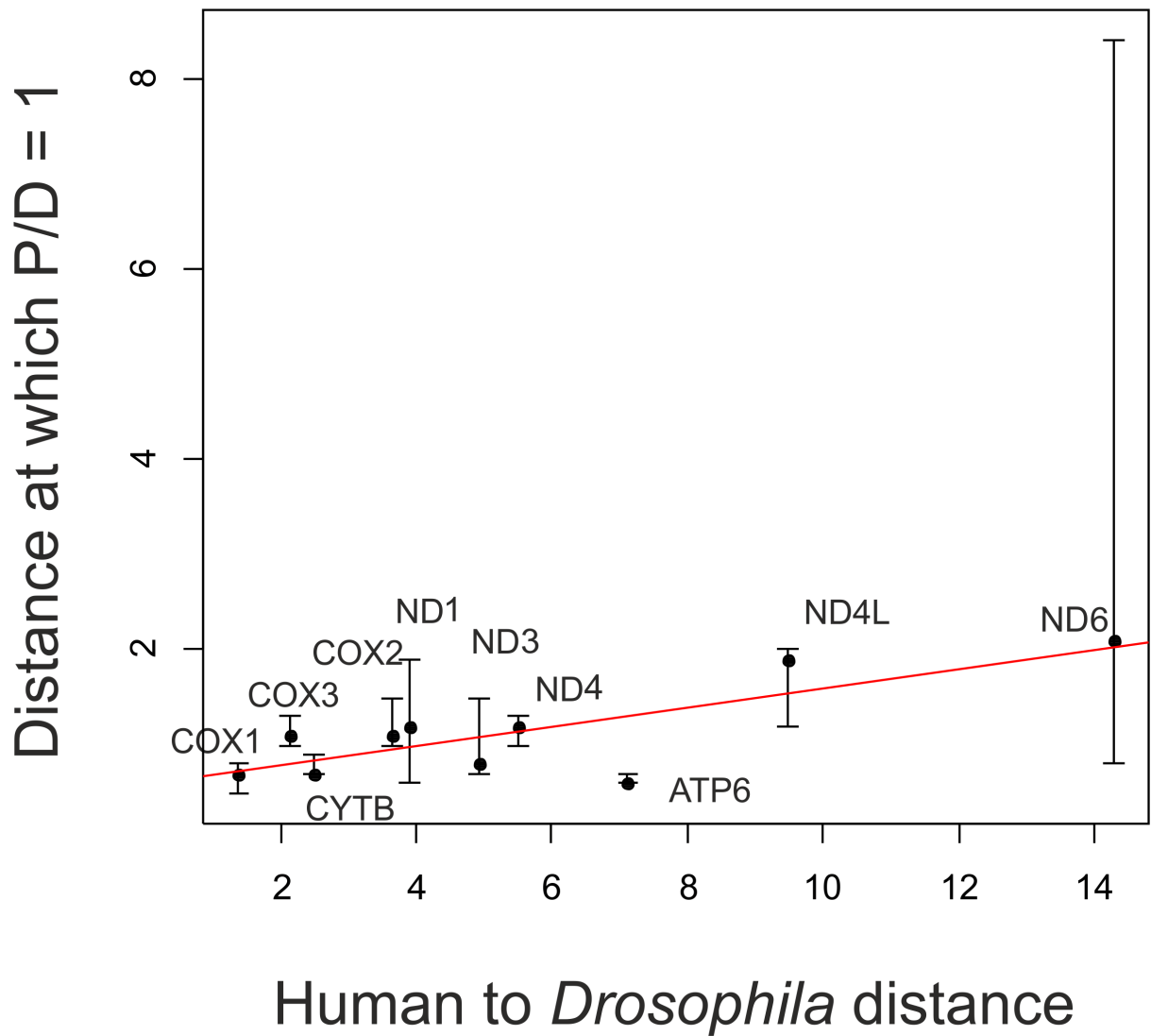
6

7

8

9

10



1  
2  
3  
4  
5  
6  
7  
8

