bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Article; Discoveries section

Pathogenic amino acid variants in mitochondrial proteins more frequently arise in lineages closely related to human than in distantly related lineages

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

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

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

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

Abstract

Propensities for different amino acids within a protein site change in the course of evolution, so that an amino acid deleterious in a particular species may be acceptable at the same site in a different species. Here, we study the amino acid-changing variants in human mitochondrial genes, and analyze their occurrence in non-human species. By reconstructing the phylogeny of mitochondrial proteins from several thousand opisthokonts, we study the changes in the rate at which either the human or the non-human allele originated, and thus determine how the fitness conferred by different human alleles changes with evolutionary distance from the human. Substitutions giving rise to the amino acid that is fixed at this site in humans, and to the non-reference allele at human polymorphic sites, tend to occur in lineages closely related to human more frequently than in more distantly related lineages, indicating that a human variant is more likely to be deleterious in more distant species. Unexpectedly, amino acids corresponding to pathogenic alleles in humans also more frequently originate at more closely related lineages, indicating that they, too, become even more deleterious with increased phylogenetic distance from human. The pathogenic variants are more similar to the human reference variant in their physico-chemical properties than a random amino acid occurring at the site in other species. Therefore, a pathogenic variant still tends to be more acceptable in human mitochondria than a variant that may only be observed after a substantial perturbation of the protein structure that occurs over long evolutionary times.

Introduction

Fitness conferred by a particular allele depends on a multitude of factors, both internal to the organism and external to it. Therefore, relative preferences for different alleles change in the course of evolution due to changes in interacting loci or in the environment. In particular, changes in the propensities for different amino acid residues at a particular protein position, or single-position fitness landscape, have been detected using multiple approaches (SPFL; Bazykin 2015; Harpak et al. 2016; Storz 2016).

bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

One way to observe such changes is by analyzing how amino acid variants (alleles) and substitutions giving rise to them are distributed over the phylogenetic tree. In particular, multiple substitutions giving rise to the same allele, or homoplasies, are more frequent in closely related than in distantly related species – a pattern expected if frequent homoplasies mark the segments of the phylogenetic tree where the arising allele confers high fitness (Rogozin et al. 2008; Povolotskaya and Kondrashov 2010; Naumenko et al. 2012; Goldstein et al. 2015; Zou and Zhang 2015; Klink and Bazykin 2017 in print).

Another manifestation of changes in SPFL is the fact that amino acids deleterious and, in particular, pathogenic in humans are often fixed as the wild type in other species. This phenomenon has been termed compensated pathogenic deviations, under the assumption that human pathogenic variants are non-pathogenic in other species due to compensatory (or permissive) changes elsewhere in the genome (Kondrashov et al. 2002; Soylemez and Kondrashov 2012; Jordan et al. 2015). The distribution of the evolutionary distances to the nearest vertebrate species in which a human pathogenic variant is observed in the nuclear genome is well approximated by a sum of two exponential distributions, suggesting that just one compensatory change is typically required to permit a formerly deleterious variant (Jordan et al. 2015).

SPFLs of mitochondrial protein-coding genes also change with time (Goldstein et al. 2015; Zou and Zhang 2015; Klink and Bazykin 2017 in print), and some of these changes may be due to intragenic or intergenic epistasis (Ji et al. 2014; Xie et al. 2016). Here, we use our previously developed approach for the study of phylogenetic clustering of homoplasies at individual protein sites (Klink and Bazykin 2017 in print) to ask how the fitness conferred by the amino acid variants observed in human mitochondrial proteins, either as benign or damaging, changes with phylogenetic distance from the human.

Results

Substitutions giving rise to reference human amino acids are more frequent in species closely related to human

The phylogenetic distribution of homoplasic (parallel, convergent or reversing) substitutions is relevant for understanding which amino acids are permitted at a particular species, and which are selected against. In particular, an excess of such substitutions giving rise to a particular allele in closely related lineages implies that the fitness conferred by this allele in these species is higher than in other species. Here, we reanalyzed the phylogenetic data on each of the 12 mitochondrial protein-coding genes in several thousand species of metazoans (Klink and Bazykin 2017 in print), as well as a set of 5 mitochondrial protein-coding genes in 4350 species of opisthokonts, focusing on the amino acids present in the human mitochondrial genome. We asked how homoplasic substitutions giving rise to the human variant are positioned phylogenetically relative to the human branch, compared with other (divergent) substitutions.

bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The reference human allele is also observed, on average, in over half of other considered species (60% of all species of metazoans for the 12-genes dataset, and 71.2% of species of opisthokonts for the 5-genes dataset). In 10% (7.1%) of these species, it did not share common ancestry with the human, but instead originated independently in an average of 21 (30) independent homoplasic substitutions per site (supplementary table 1, supplementary figs. 1-2, Supplementary Materials online).

We asked whether the phylogenetic positions of homoplasic substitutions giving rise to the human allele are biased, compared to the positions of divergent substitutions giving rise to other alleles. This analysis controls for the biases associated with pooling sites and amino acid variants (see Materials and Methods).

In most proteins, the mean phylogenetic distances from human to the branches at which the human reference amino acid emerged independently due to a homoplasy were ~10% shorter than to the branches at which another amino acid emerged (fig. 1, fig. 3a). No such decrease was observed for a random amino acid among those that were observed at this site in non-human species, or in simulated data (fig. 1, fig. 3a). The number of homplasic substitutions giving rise to the human reference amino acid relative to the number of divergent substitutions towards non-human amino acids (H/D ratio) uniformly decreases with the evolutionary distance from the human (fig. 2, fig. 3b). At most genes, the relative number of homoplasic substitutions giving rise to the human allele drops 1.5-3-fold with phylogenetic distance from the human branch (fig. 2, fig. 3b).

Figure 1

Figure 2

Figure 3

The excess of homoplasic changes to the human reference amino acid at small phylogenetic distances from human is not an artefact of differences in mutation rates between amino acids in distinct clades, since these rates are similar in all considered species and cannot lead to such clustering (Klink and Bazykin 2007 in print). It is also not an artefact of differences in codon usage bias between species, as it was still observed when we considered only “accessible” amino acid pairs where the derived amino acid could be reached through a single nucleotide substitution from any codon of the ancestral amino acid (fig. 1).

Amino acids corresponding to variant alleles at human polymorphic sites more frequently arise in species closely related to humans

Next, we considered human SNPs in mitochondrial proteins in the MITOMAP database (Lott et al. 2013). We analyzed the phylogenetic distribution of substitutions in non-human species giving rise to the amino acid that is also observed as the non-reference (usually minor) allele in humans (supplementary table 2, Supplementary Materials online).

bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Similarly to the human reference amino acid variant, the homoplasic substitutions giving rise to non-reference alleles were clustered on the phylogeny near humans, compared to the divergent substitutions giving rise to a variant never observed in humans (fig. 4-6). Again, the mean phylogenetic distance from human to a substitution giving rise to the human non-reference amino acid was ~10% lower than to other substitutions (fig. 4, fig. 6a), and the density of homoplasic substitutions giving rise to such an allele dropped significantly with phylogenetic distance from the human branch (fig. 5, Fig. 6b). As before, this clustering was also observed if only accessible pairs of amino acids were considered, while no systematic differences were observed for random amino acids (fig. 4).

Figure 4

Figure 5

Figure 6

Amino acids corresponding to human pathogenic variants more frequently arise in species closely related to humans

Finally, we considered human alleles annotated as disease-causing in the MITOMAP database (Lott et al. 2013). Since only a handful of mutations is thus annotated in each gene (supplementary table 3, Supplementary Materials online), the variance in the estimates of the H/D ratio is, as expected, large. Still, in five of the twelve genes of the metazoan dataset (fig. 7-8), as well as in the opisthokont dataset (fig. 9), the human pathogenic variant also arose independently more frequently in the phylogenetic vicinity of humans. The opposite pattern, i.e., biased occurrence of the human pathogenic variant in phylogenetically remote species, has not been observed in any of the genes. This trend was even stronger for the six pathogenic mutations confirmed by two or more independent studies ("confirmed" status in MITOMAP; fig. 9b). As before, this result is not due to preferential usage of codons more likely to mutate into the human variant in species closely related to human (fig. 7).

Figure 7

Figure 8

Figure 9

Human pathogenic variants are more biochemically similar than non-human variants to normal human variants

To understand what drives the preferential emergence of the human variant, either normal or pathogenic, in species closely related to humans, we analyzed the identity of these variants. Both normal (Fig. 10a) and pathogenic (Fig. 10b) human amino acids were more similar in their biochemical properties according to the Miyata matrix (Miyata et al. 1979) to the human reference variant than amino acids observed in non-human species. In turn, amino acids observed in non-human

bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

species were more similar to the human reference variant than amino acids never observed at this site in any species.

Figure 10.

Individual mutations

To illustrate the observed phylogenetic clustering, we plotted the distribution over the opisthokont phylogeny of substitutions at the 6 amino acid sites that carry pathogenic mutations with "confirmed" status (supplementary fig. 3, Supplementary Materials online). Visual inspection of these plots confirms that the substitutions giving rise to pathogenic alleles tend to be clustered in the vicinity of the human, compared to other substitutions of the same ancestral amino acids (supplementary fig. 3, Supplementary Materials online). For the metazoan dataset, we also plot three select individual amino acid sites with known pathogenic mutations which are considered below.

The V→A mutation at ND1 site 113 has been reported to cause bipolar disorder, and decreases the mitochondrial membrane potential and reduces ND1 activity in experiments (Munakata et al. 2004). According to the mtDB database (Ingman and Gyllensten 2006), the A allele persists in human population at 0.5% frequency. However, we observed that the same allele originated independently in three clades of vertebrates: Old World monkeys (Cercopithecidae), flying lemurs (Cynocephalidae) and turtles (Geoemydidae), while most of the other substitutions of V at this site occurred in invertebrates (fig. 11). As a result, the mean phylogenetic distance between human and the parallel V→A substitutions is 2.35 (median 0.75), while it is 4.12 (median 4.6) for substitutions of V to other amino acids.

Figure 11

The V→A mutation at COX3 site 91 has been reported to cause Leigh disease (Mkaouar-Rebai et al. 2011). In metazoans, A allele at this site has originated independently 7 times, including 6 times from V and once from I. All but one of these substitutions occurred in mammals, while tens of substitutions of V and I resulting in other amino acids occurred throughout metazoans (Fig. 12). As a result, the mean phylogenetic distance from human to V→A substitutions was 0.6 (median 0.7), while the distance to other mutations from V was 1.8 (median 2.1); the corresponding numbers for I were 0.6 (median 0.6) and 2.7 (median 2.2).

Figure 12

The I→V mutation at ND6 site 33 has been reported to cause type two diabetes (Tawata et al. 2000) and has population frequency of 0.1% according to the mtDB. However, this substitution has occurred repeatedly in parallel in mammals and amphibians, while other substitutions of I were frequent in invertebrates (Fig. 13). The mean phylogenetic distance from human to parallel substitutions to V was 2.4 (median 1.5), while it was as high as 12.9 (median 14.8) for other amino acids.

bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 13

Discussion

A variant deleterious in human may be fixed in a non-human species, and sometimes this can be explained by compensatory or permissive mutations elsewhere in the genome (Kondrashov et al. 2002, Kern and Kondrashov 2004, Jordan et al. 2015). Here, we reveal the opposite facet of the same phenomenon: a variant that is fixed or polymorphic in human may be deleterious in a non-human species.

Indeed, we find that substitutions giving rise to the human allele occur in species that are more closely related to *H. sapiens* than species in which substitutions to other amino acid occur. While artefactual evidence for excess of parallel substitutions between closely related species may arise from discordance between gene trees and species trees (Mendes et al. 2016), it is unlikely that it causes the observed signal in our analysis. For reference alleles, we have previously shown that phylogenetic clustering in mitochondrial proteins is not due to tree reconstruction errors (Klink and Bazykin 2017 in print), and mitochondrial genomes do not recombine, which makes other causes of discordance unlikely. For variants polymorphic in humans, artefactual evidence for parallelism could theoretically also arise from a variant that was polymorphic in the last common ancestor of human and another species such as chimpanzee, was subsequently fixed in this other species, and survived as polymorphism in the human lineage until today. However, the last common ancestor of human mitochondrial lineages probably lived no longer than 148 thousand years ago (Poznik et al. 2013), which is much more recent than the time of human-chimpanzee divergence, also excluding this option.

Therefore, the observed phylogenetic clustering implies a decrease of the fitness conferred by the human amino acid relative to that conferred by other amino acids with phylogenetic distance from human.

Arguably, one would expect the opposite pattern in the variants pathogenic for humans. Indeed, it is likely that the majority of such variants are also deleterious in species related to humans, while changes in the genomic context in more distantly related species may make these variants tolerable.

Instead, we observe the pattern similar to that for the non-pathogenic variants: the amino acid variants pathogenic for humans are also relatively more likely to emerge independently as a homoplasy in species closely related to *H. sapiens* than in more distantly related species.

The similarity between the phylogenetic distribution of the benign and pathogenic variants could be due to some of the benign mutations being incorrectly annotated as pathogenic (Exome Aggregation Consortium et al. 2015). However, none of the six mutations with “confirmed” status in Mitomap are present in the mtDB database among the 2704 sequences from different human

bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

populations, suggesting that these mutations are indeed damaging, while they demonstrate a pronounced clustering (fig. 9).

Consideration of biochemical similarities of amino acid variants helps explain why human-pathogenic variants can still be more likely to occur in species closely related to humans. We find that despite their pathogenicity, the human pathogenic variants are on average more biochemically similar to the major human allele than other amino acids that were observed at this site in non-human species. Therefore, in the context of the human genome, the annotated human pathogenic variant probably disrupts the protein structure less than an "alien" non-human variant. Conversely, many alien variants that are not observed in humans are likely to be even more deleterious in human than the annotated pathogenic allele, perhaps lethal, while they confer high fitness in the genomic context of their own species.

In summary, we have shown that all types of alleles, including pathogenic, that occur in human mitochondrial protein-coding genes more frequently emerge independently in species more closely related to *H. sapiens*. Such a decrease in occurrence of human amino acids with phylogenetic distance from human is probably due to a higher similarity of the sequence and/or environmental context in more closely related species. More generally, it is broadly accepted that the occurrence of a mutation in another species is an important predictor of its pathogenicity in humans (Adzhubei et al. 2010; Kumar et al. 2011), and it is increasingly appreciated that it is important to account for the degree of relatedness of the considered species to human (Jordan et al. 2015). Our observation that human-pathogenic alleles are underrepresented in more distantly related, rather than in more closely related, species shows that the direction of the association between relatedness and prediction of pathogenicity can be counterintuitive.

Materials and Methods

Data

Alignments of 12 mitochondrial proteins of metazoans (Breen et al. 2012) and a joint alignment of five concatenated mitochondrial genes of opisthokonts were obtained as described in (Klink and Bazykin 2017 in print). These alignments were used to reconstruct constrained phylogenetic trees, ancestral states and phylogenetic positions of substitutions (Klink and Bazykin 2017 in print). As the reference human allele, we used the revised Cambridge Reference Sequence (rCRS) of the human mitochondrial DNA (Andrews et al. 1999). As non-reference alleles, we used amino acid changing variants from "mtDNA Coding Region & RNA Sequence Variants" section of the MITOMAP database. As pathogenic alleles, we used amino acid changing variants with "reported" (i.e., supported by one publication) or "confirmed" (i.e., supported by at least two independent publications) status from the "Reported Mitochondrial DNA Base Substitution Diseases: Coding and Control Region Point Mutations" section of MITOMAP.

Clustering of substitutions giving rise to the human amino acid around the human branch

bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

For each amino acid site in a protein, we considered those amino acid variants that (i) constitute the reference allele in humans and had arisen in the human lineage at some point during its evolution, or (ii) had originated in humans as a derived polymorphic allele, or (iii) are annotated in humans as pathogenic alleles. For further consideration, we retained only such alleles from each class for which at least one homoplasic (i.e., giving rise to the same allele by way of parallelism, convergence, or reversal) and at least one divergent (i.e., giving rise to a different allele) substitution from the same ancestral variant was observed at this site elsewhere on the phylogeny outside of the human lineage. Substitutions, including reversals, that occurred anywhere on the path between the root and *H. sapiens* were excluded. While the homoplasic and divergent substitutions had to derive from the same ancestral variant, it could be either the same or a different variant than that ancestral to the variant observed in human.

For each such allele, we compared the phylogenetic distances between human and positions of homoplasic substitutions with the distances between human and positions of divergent substitutions, using a previously described procedure which controls for the differences in SPFLs between sites or in mutational probabilities of different substitutions (Klink and Bazykin 2017 in print). Briefly, for each ancestral amino acid at each site, we subsampled equal numbers of homoplasic substitutions to human (reference, non-reference or pathogenic) amino acids and divergent substitutions to non-human amino acids. We then pooled these values across all considered sites and groups of derived alleles. Next, we categorized them by the phylogenetic distance between the human and the position of the substitution, and calculated, for each bin of the phylogenetic distances, the ratio of the numbers of homoplasic (H) and divergent (D) substitutions (H/D). To obtain the mean values and 95% confidence intervals for the H/D statistic, we bootstrapped sites in 1000 replicates, each time repeating the entire resampling procedure. As a control, we performed the same analyses using instead of the human variant a random non-human amino acid among those observed at this site, or using data obtained by simulating the evolution at each site along the same phylogeny and with gene-specific GTR+Gamma amino acid substitution matrices (Klink and Bazykin 2017 in print).

Acknowledgements

This work was supported by the Russian Science Foundation (grant number 14-50-00150).

We thank Shamil Sunyaev, Alexey Kondrashov, Dmitry Pervouchine and Vladimir Seplyarskiy for valuable comments.

References.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat. Methods 7:248–249.

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat. Genet. 23:147.

bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

Exome Aggregation Consortium, Monkol Lek, Konrad Karczewski, Eric Minikel, Kaitlin Samocha, Eric Banks, Timothy Fennell. 2015. Analysis of protein-coding genetic variation in 60,706 humans. Available from: http://biorxiv.org/lookup/doi/10.1101/030338

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

Harpak A, Bhaskar A, Pritchard JK. 2016. Effects of variable mutation rates and epistasis on the distribution of allele frequencies in humans. Available from: http://biorxiv.org/lookup/doi/10.1101/048421

Ingman M, Gyllensten U. 2006. mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. Nucleic Acids Res. 34:D749–D751.

Ji Y, Liang M, Zhang J, Zhang M, Zhu J, Meng X, Zhang S, Gao M, Zhao F, Wei Q-P, et al. 2014. Mitochondrial haplotypes may modulate the phenotypic manifestation of the LHON-associated ND1 G3460A mutation in Chinese families. J. Hum. Genet. 59:134–140.

Jordan DM, Frangakis SG, Golzio C, Cassa CA, Kurtzberg J, Task Force for Neonatal Genomics, Davis EE, Sunyaev SR, Katsanis N. 2015. Identification of cis-suppression of human disease mutations by comparative genomics. Nature 524:225–229.

Klink GV, Bazykin GA. 2017. Parallel evolution of metazoan mitochondrial proteins. GBE in print.

Kondrashov AS, Sunyaev S, Kondrashov FA. 2002. Dobzhansky-Muller incompatibilities in protein evolution. Proc. Natl. Acad. Sci. U. S. A. 99:14878–14883.

Kumar S, Dudley JT, Filipski A, Liu L. 2011. Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. Trends Genet. TIG 27:377–386.

Lott MT, Leipzig JN, Derbeneva O, Xie HM, Chalkia D, Sarmady M, Procaccio V, Wallace DC. 2013. mtDNA Variation and Analysis Using MITOMAP and MITOMASTER. Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al 1:1.23.1–1.23.26.

Mendes FK, Hahn Y, Hahn MW. 2016. Gene tree discordance can generate patterns of diminishing convergence over time. Mol. Biol. Evol.

Miyata T, Miyazawa S, Yasunaga T. 1979. Two types of amino acid substitutions in protein evolution. J. Mol. Evol. 12:219–236.

bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Mkaouar-Rebai E, Ellouze E, Chamkha I, Kammoun F, Triki C, Fakhfakh F. 2011. Molecular-clinical correlation in a family with a novel heteroplasmic Leigh syndrome missense mutation in the mitochondrial cytochrome c oxidase III gene. J. Child Neurol. 26:12–20.

Munakata K, Tanaka M, Mori K, Washizuka S, Yoneda M, Tajima O, Akiyama T, Nanko S, Kunugi H, Tadokoro K, et al. 2004. Mitochondrial DNA 3644T-->C mutation associated with bipolar disorder. Genomics 84:1041–1050.

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.

Poznik GD, Henn BM, Yee M-C, Sliwerska E, Euskirchen GM, Lin AA, Snyder M, Quintana-Murci L, Kidd JM, Underhill PA, et al. 2013. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. Science 341:562–565.

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.

Soylemez O, Kondrashov FA. 2012. Estimating the rate of irreversibility in protein evolution. Genome Biol. Evol. 4:1213–1222.

Storz JF. 2016. Causes of molecular convergence and parallelism in protein evolution. Nat. Rev. Genet. 17:239–250.

Tawata M, Hayashi JI, Isobe K, Ohkubo E, Ohtaka M, Chen J, Aida K, Onaya T. 2000. A new mitochondrial DNA mutation at 14577 T/C is probably a major pathogenic mutation for maternally inherited type 2 diabetes. Diabetes 49:1269–1272.

Xie S, Zhang J, Sun J, Zhang M, Zhao F, Wei Q-P, Tong Y, Liu X, Zhou X, Jiang P, et al. 2016. Mitochondrial haplogroup D4j specific variant m.11696G > a(MT-ND4) may increase the penetrance and expressivity of the LHON-associated m.11778G > a mutation in Chinese pedigrees. Mitochondrial DNA Part DNA Mapp. Seq. Anal.:1–8.

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.

Figure legends

bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 1. Ratios of the phylogenetic distances between the human branch and substitutions to the human reference allele vs. to other amino acids. Ratios <1 imply that the considered allele arises independently closer at the phylogeny to humans than other alleles. 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 (*, P<0.05; **, P<0.01;***, P<0.001). ref, human reference allele; access, human reference alleles from accessible amino acid pairs (see text); rand, a random non-human amino acid among those that were present in the site; sim, human allele in simulated data.

Figure 2. Higher fraction of homoplasic substitutions to the human reference amino acid, compared with random amino acids that had independently originated at this site (H/D ratio), in species closely related to human. Horizontal axis, distance between branches carrying the substitutions and the human branch, measured in numbers of amino acid substitutions per site, split into bins by $\log_2$(distance). Vertical axis, H/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 H/D ratio of 1. Arrows represent the distance between human and *Drosophila.*

Figure 3. A reduced ratio of the phylogenetic distances between the human branch and substitutions to the considered amino acid vs. to other amino acids at the same site (a) and a higher H/D ratio in species closely related to human (b) are observed for the human reference amino acid, but not for a random allele observed at this site or a simulated allele, in the 4350-species opisthokonts phylogeny. Notations same as in Figures 1 and 2.

Figure 4. Ratios of the phylogenetic distances between the human branch and substitutions to the human non-reference allele vs. to other amino acids. Notations same as in Figure 1.

Figure 5. Higher fraction of homoplasic substitutions to the human non-reference amino acid, compared with random amino acids that had independently originated at this site (H/D ratio), in species closely related to human. Notations same as in Figure 2.

Figure 6. A reduced ratio of the phylogenetic distances between the human branch and substitutions to the considered amino acid vs. to other amino acids at the same site (a) and a higher H/D ratio in species closely related to human (b) are observed for the human non-reference amino acid, but not for a random allele observed at this site or a simulated allele, in the 4350-species opisthokonts phylogeny. Notations same as in Figures 1 and 2.

Figure 7. Ratios of the phylogenetic distances between the human branch and substitutions to the human pathogenic allele vs. to other amino acids. Notations same as in Figure 1.

Figure 8. Higher fraction of homoplasic substitutions to the human pathogenic amino acid, compared with random amino acids that had independently originated at this site (H/D ratio), in species closely related to human. Notations same as in Figure 2.

bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 9. A reduced ratio of the phylogenetic distances between the human branch and substitutions to the considered amino acid vs. to other amino acids at the same site (a,b) and a higher H/D ratio in species closely related to human (c) are observed for all (a,c) and confirmed (b) human pathogenic amino acid, but not for a random allele observed at this site or a simulated allele, in the 4350-species opisthokonts phylogeny. Notations same as in Figures 1 and 2.

Figure 10. Distributions of ranks of Miyata distances between the reference human allele and the non-reference (a) or pathogenic (b) allele (white), compared with other alleles that were (gray) or were not (black) observed at the same site. For each site with known polymorphisms or pathogenic mutations, we ranked all amino acids by Miyata distance from reference human allele, and then obtained distance rank for pathogenic (or non-reference) human variant, mean rank for amino acids that occurred in a site but did not observed in human and mean rank for rest amino acids.

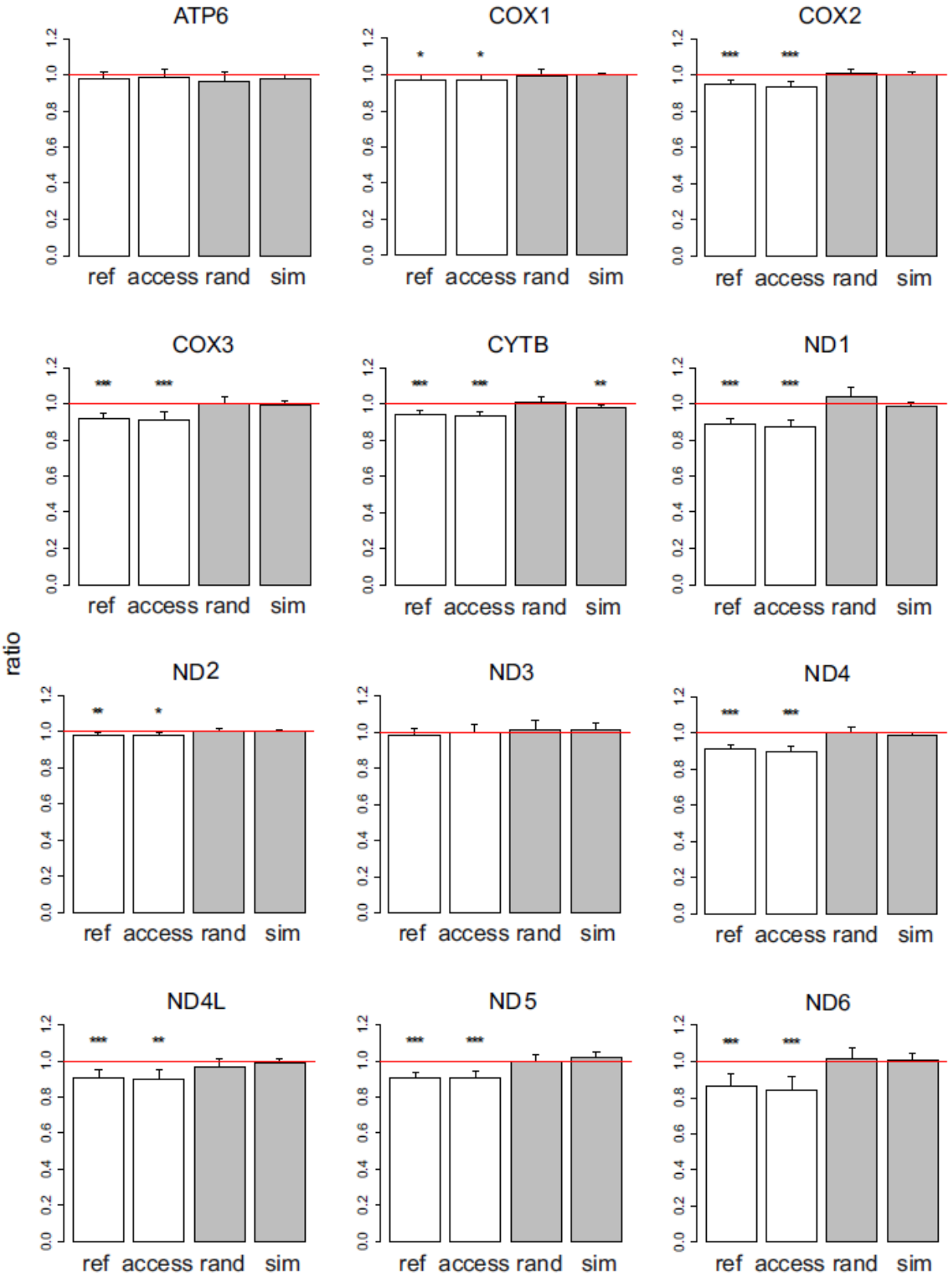
Figure 11. Substitutions in site 113 of ND1. Blue star is *H. sapiens* branch; red dots are substitutions of valine to alanine, which is pathogenic in human and dots of other colors are substitutions to other amino acids. Phylogenetic distances are measured in numbers of amino acid substitutions per site. The branches indicated with the blue waves are shortened approximately by 2 distance units.

Figure 12. Substitutions in site 91 of COX3. Blue star is *H. sapiens* branch; red dots are substitutions from valine to alanine which is pathogenic in human, black dot is substitution from isoleucine to alanine, and dots of other colors are substitutions to other amino acids. Phylogenetic distances are measured in numbers of amino acid substitutions per site.

Figure 13. Substitutions in site 33 of ND6. Blue star is *H. sapiens* branch; red dots are substitutions of isoleucine to valine, which is pathogenic in human and dots of other colors are substitutions to other amino acids. Phylogenetic distances are measured in numbers of amino acid substitutions per site. The branches indicated with the blue waves are shortened approximately by 3 distance units.

bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 1



bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure2

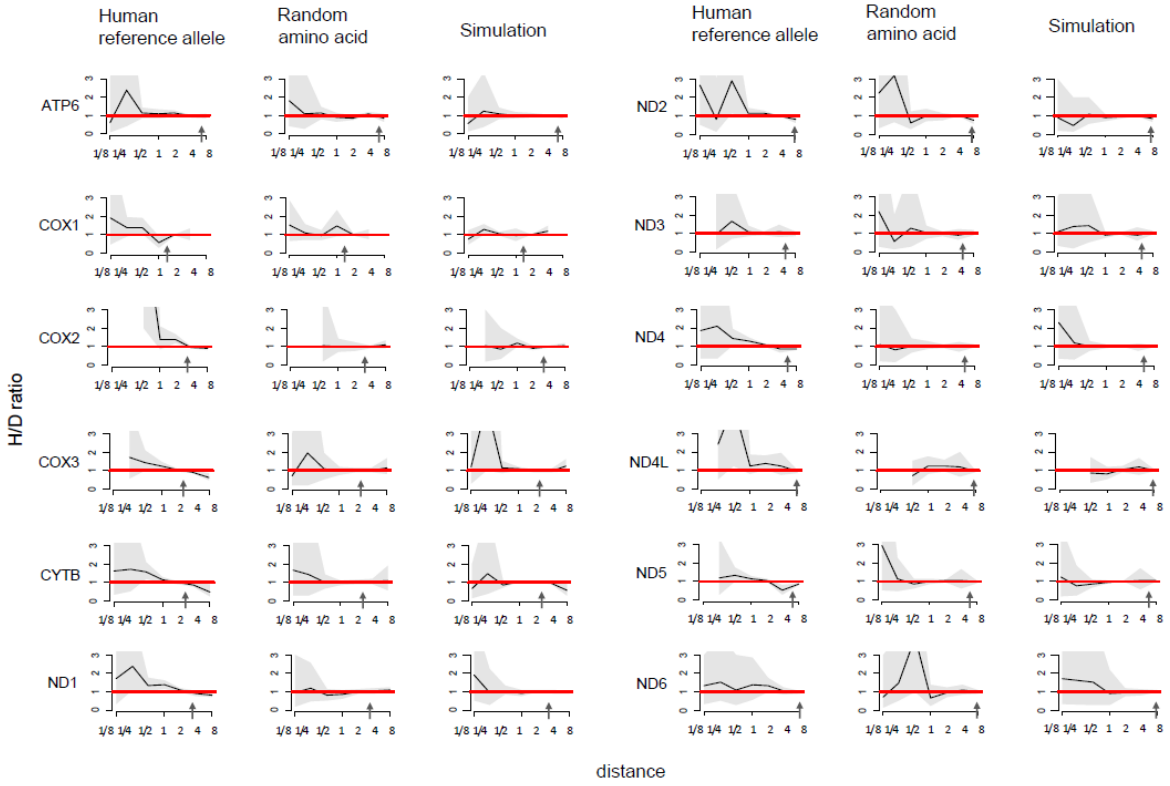
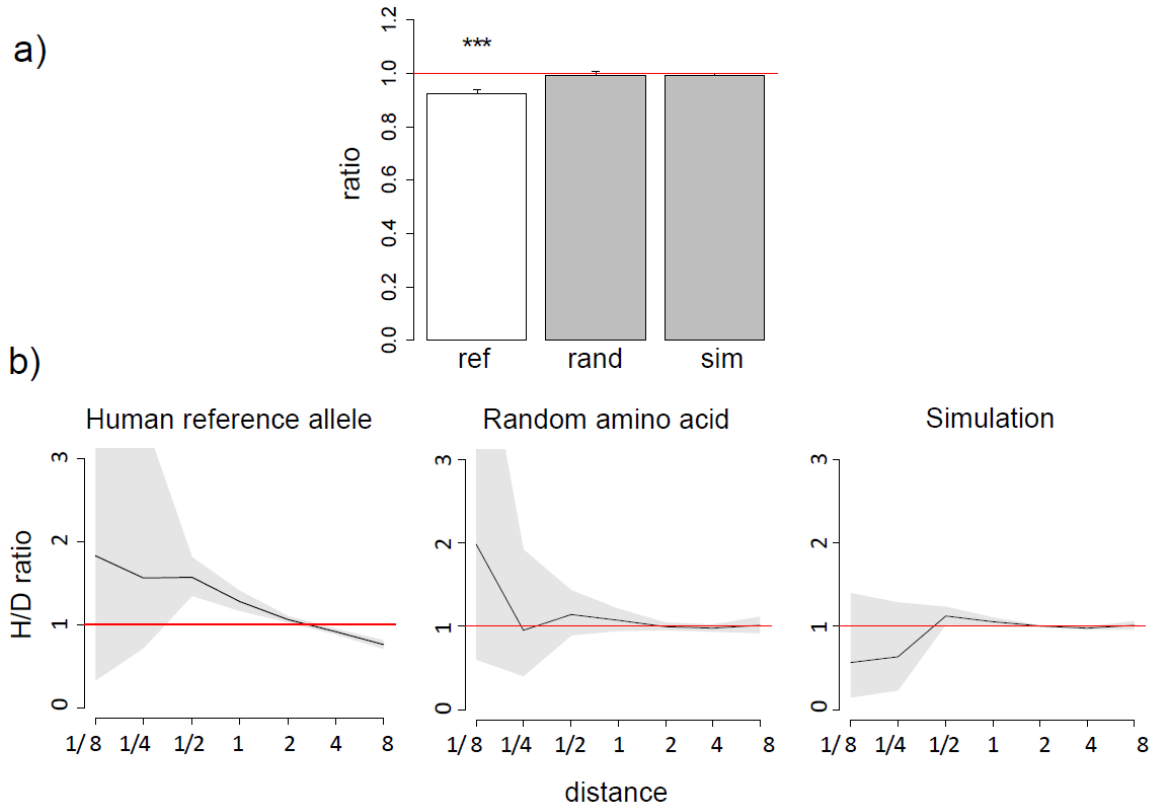
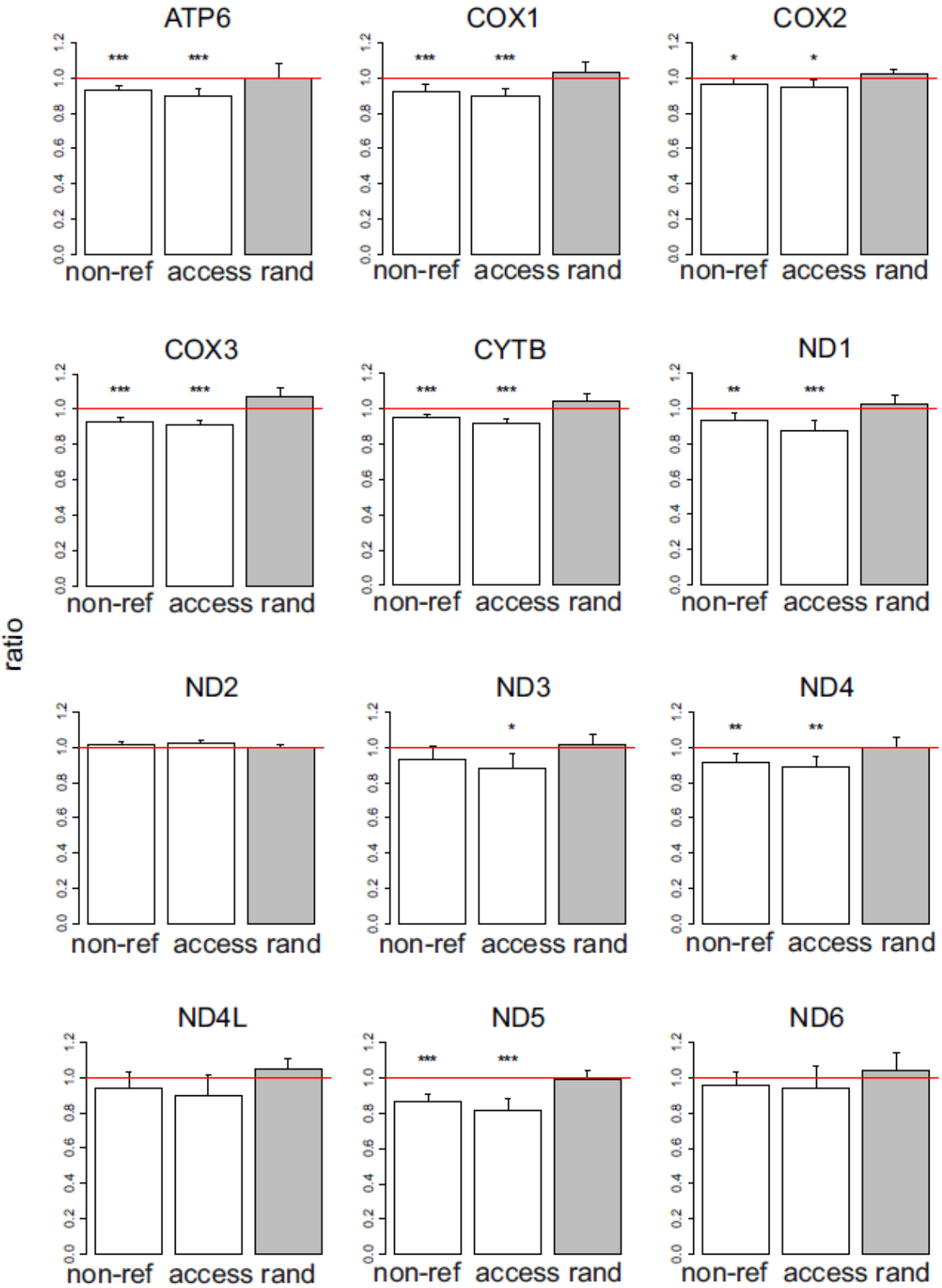


Figure 3



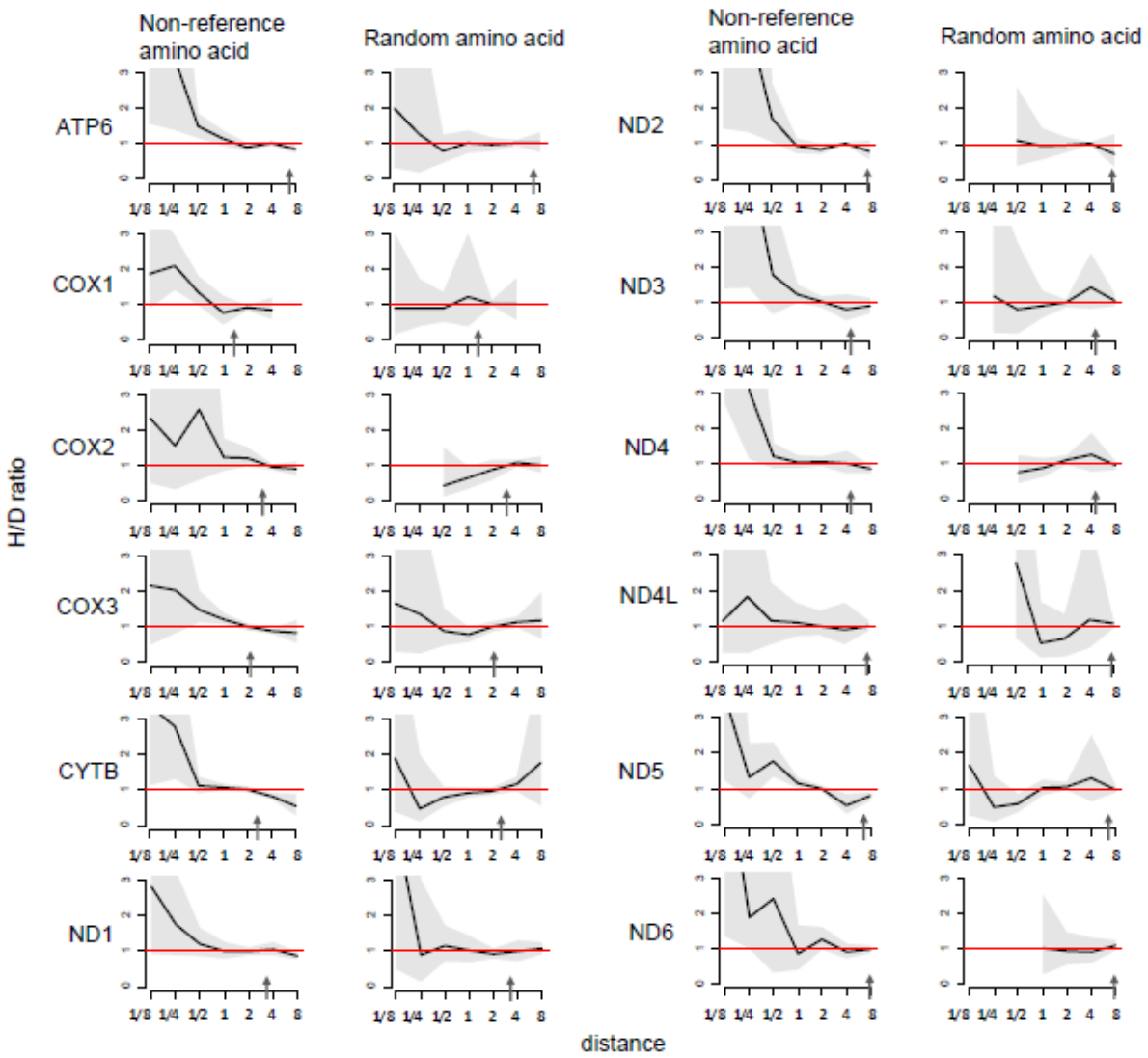
bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 4



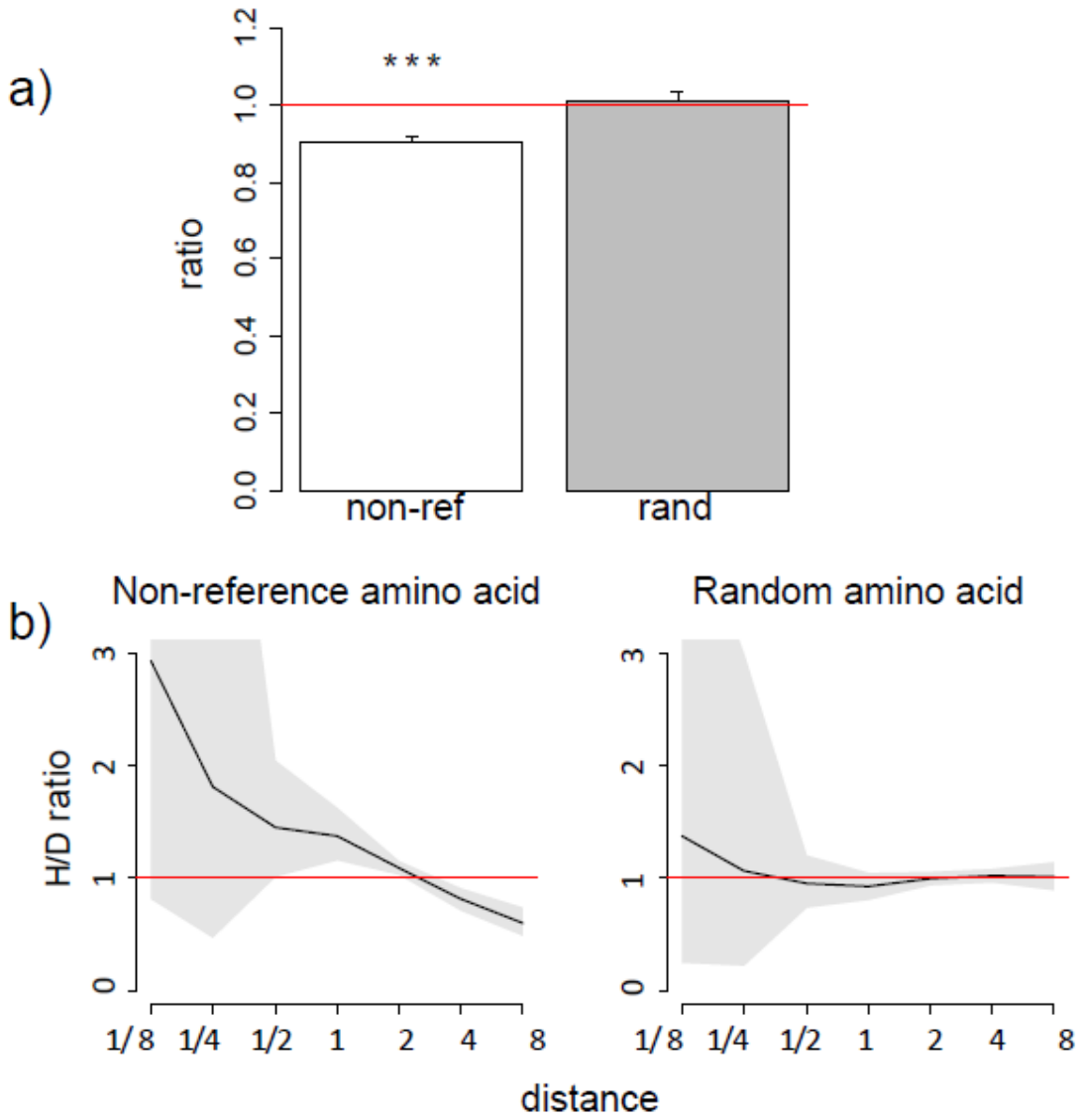
bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 5



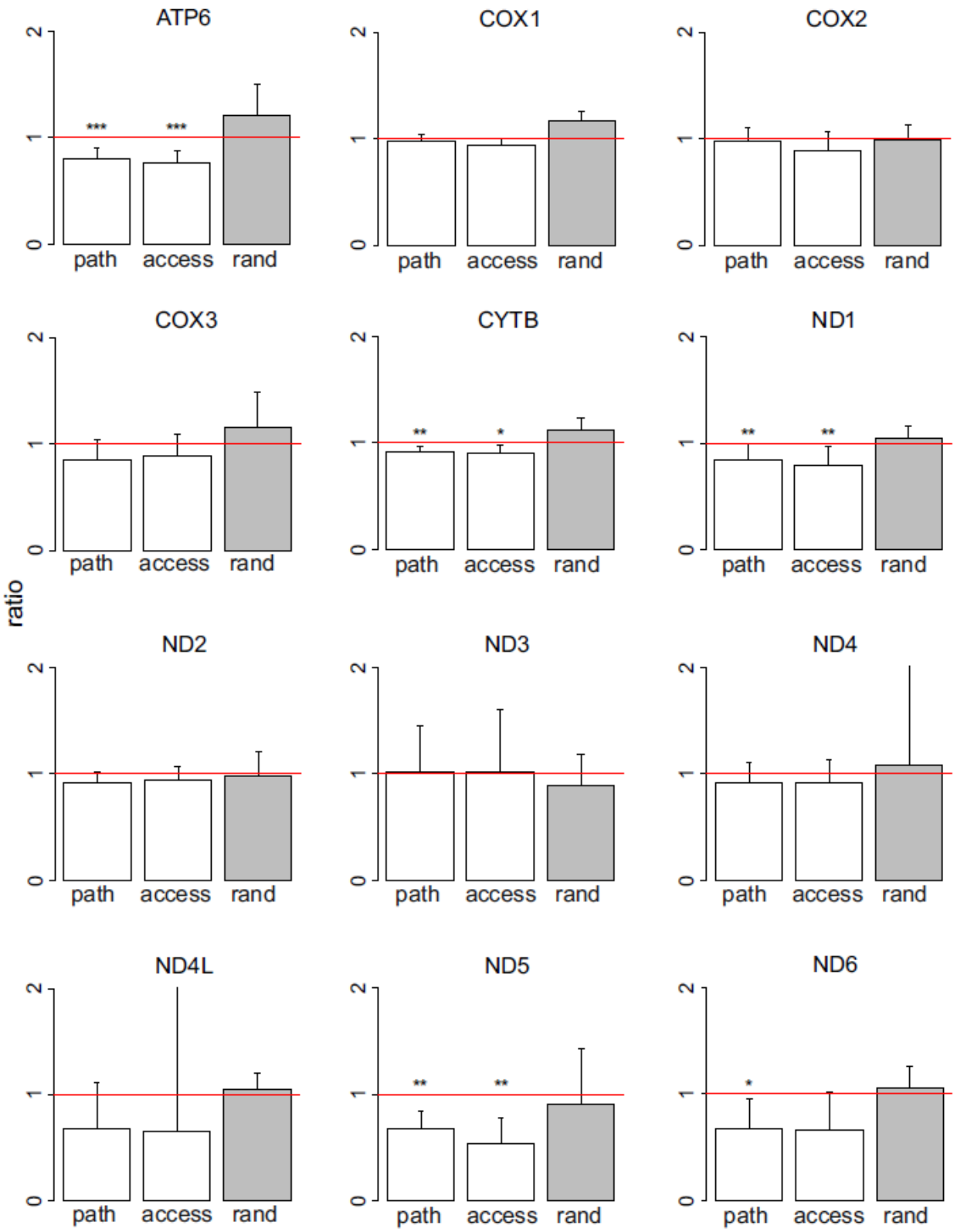
bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 6



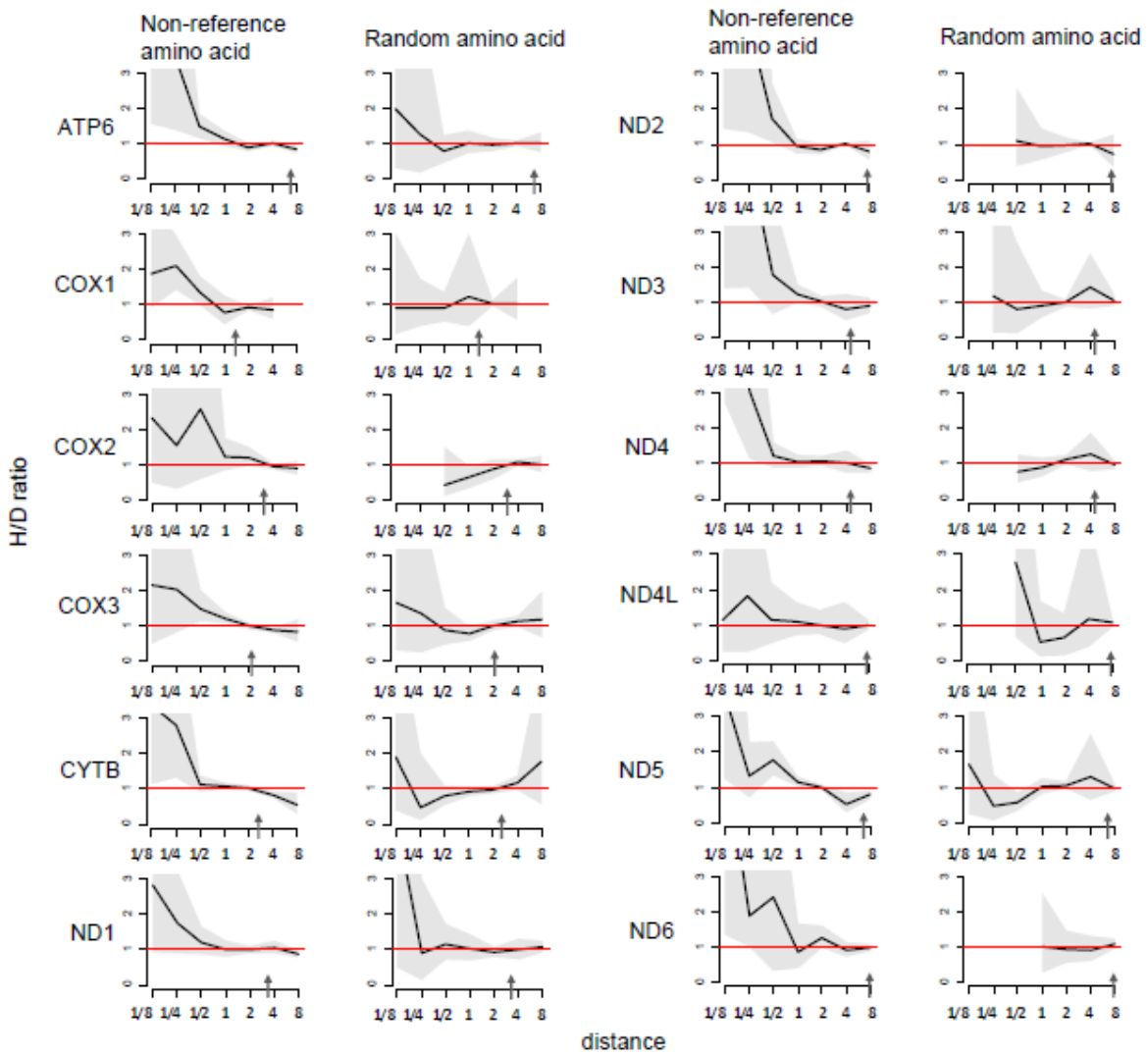
bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 7



bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 8



bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 9

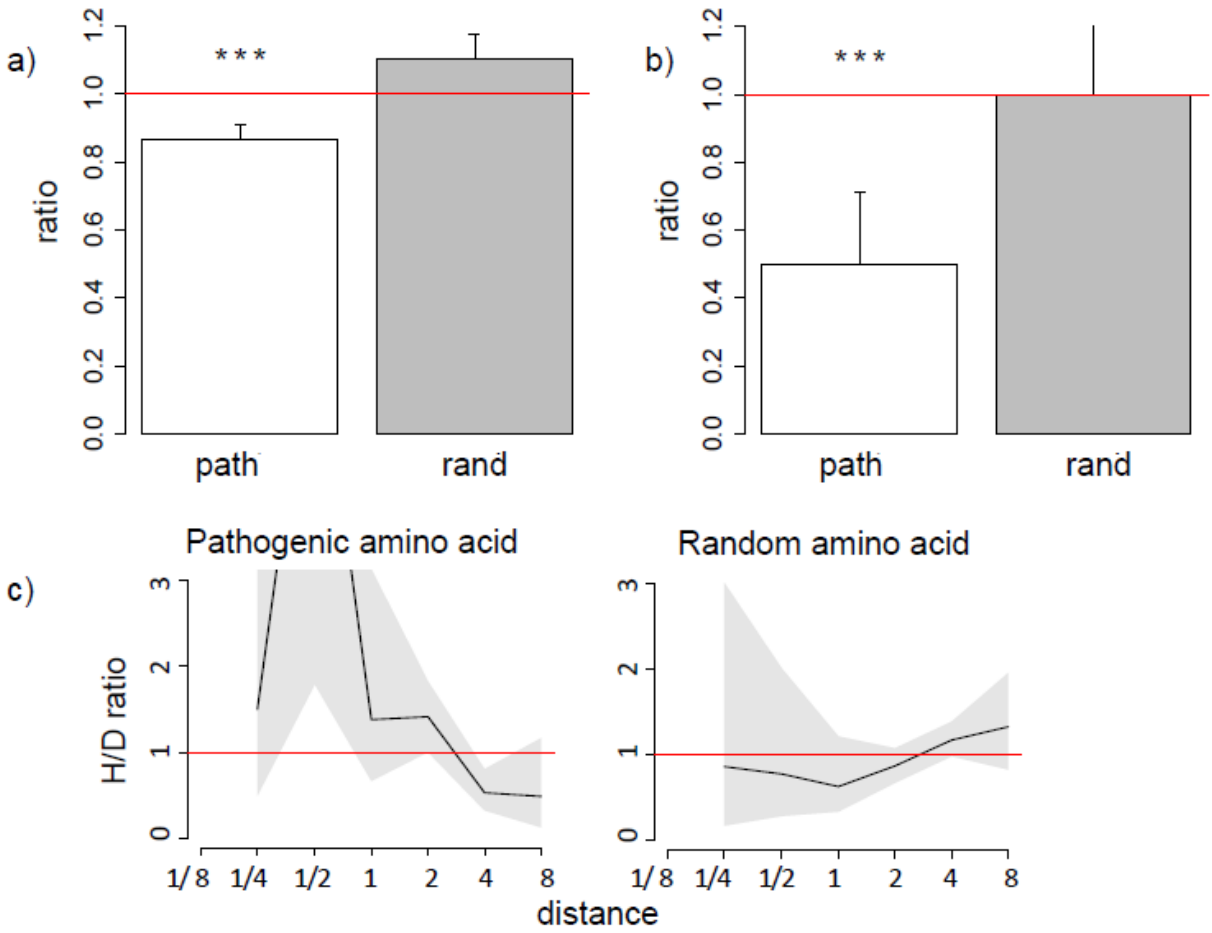
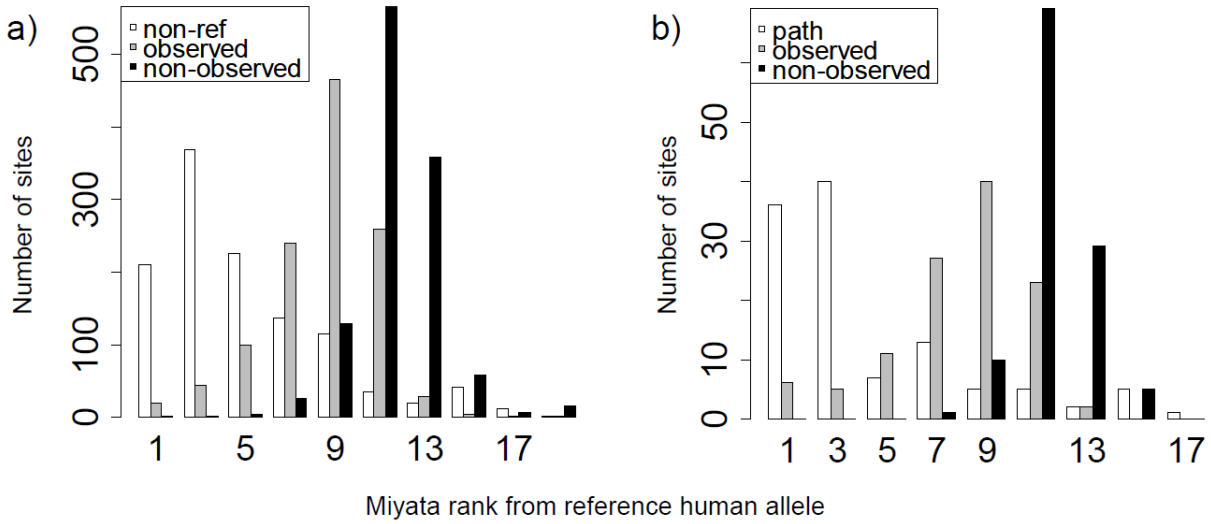
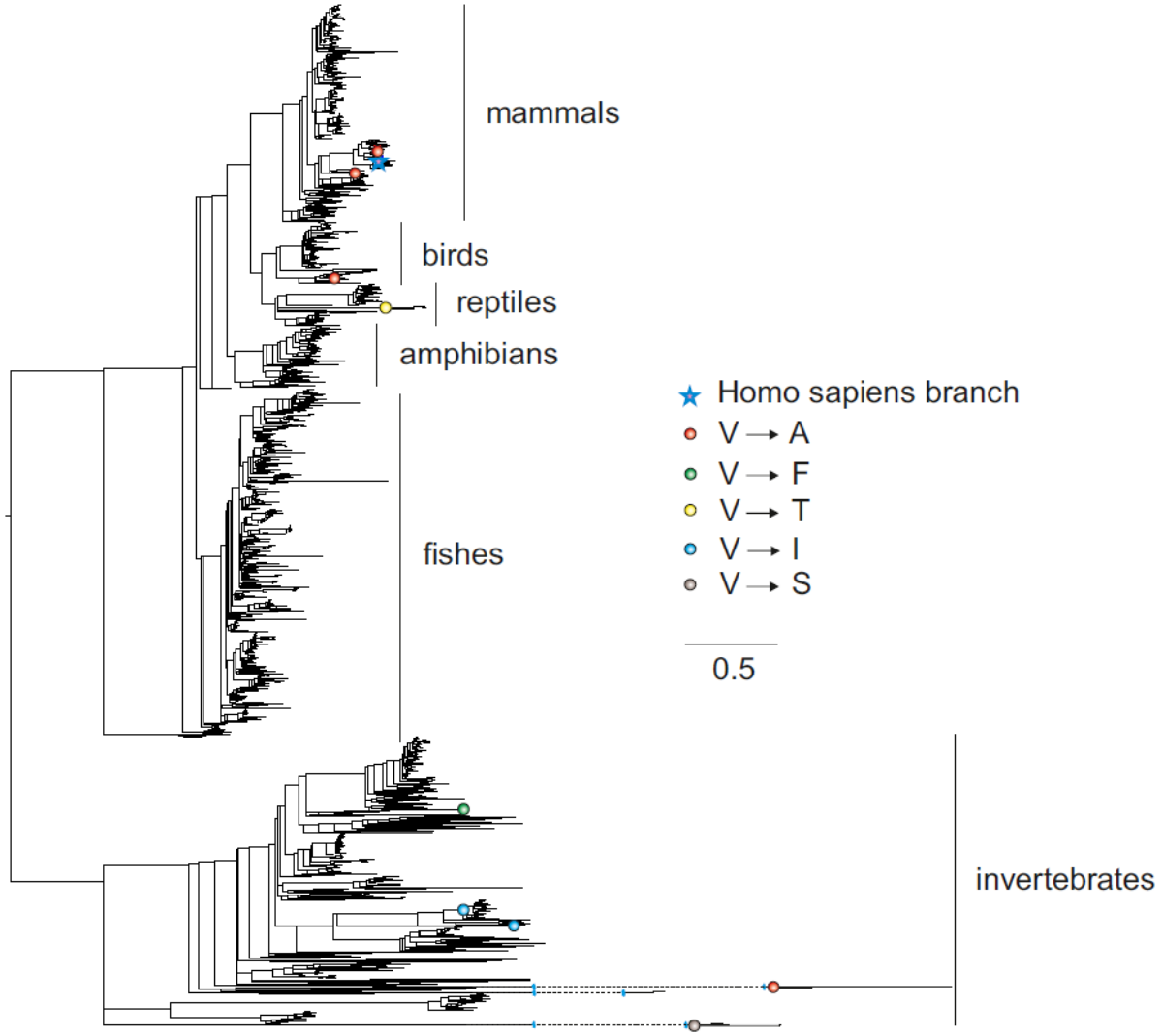


Figure 10.



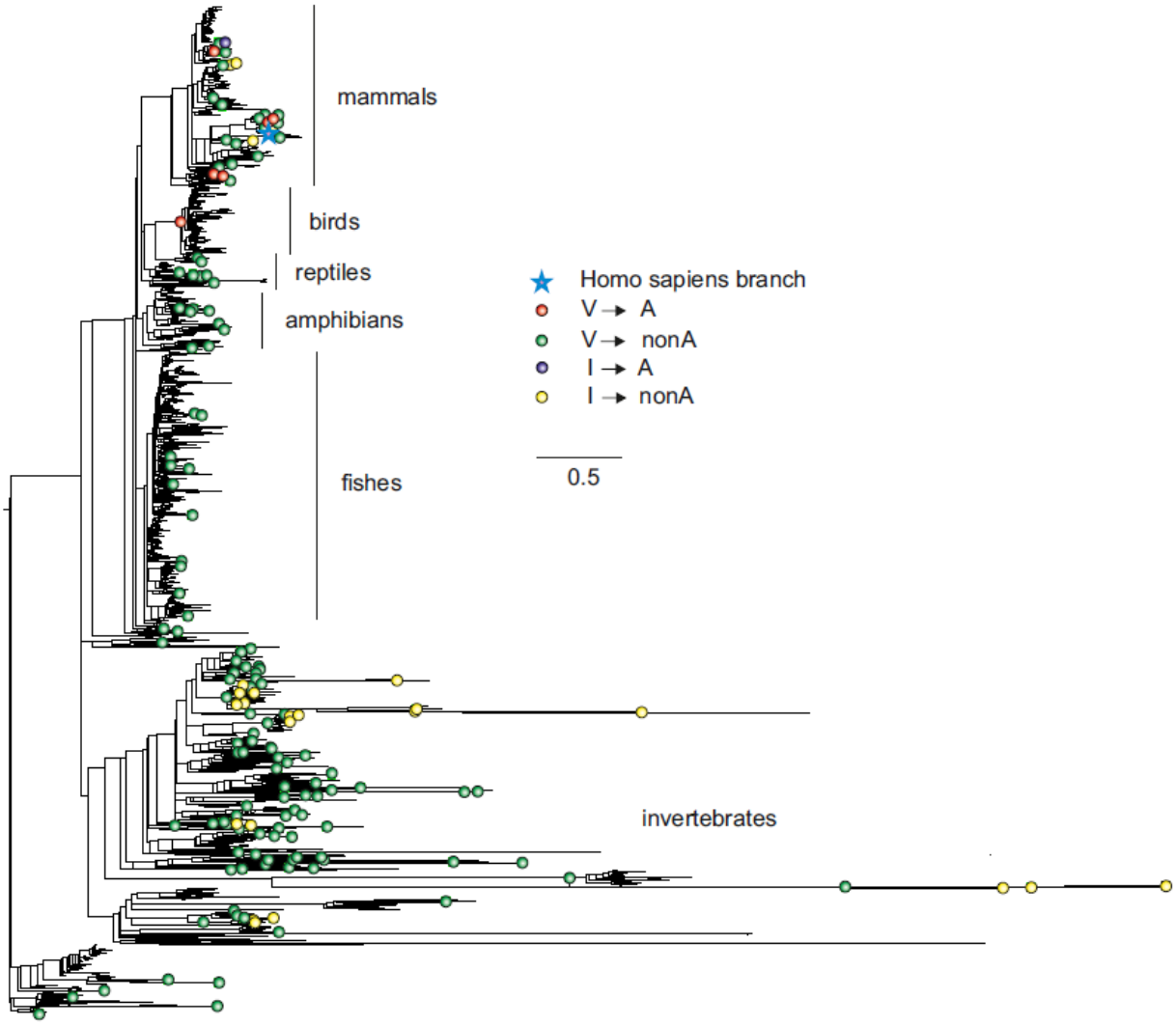
bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 11



bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 12



bioRxiv preprint doi: https://doi.org/10.1101/119792; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

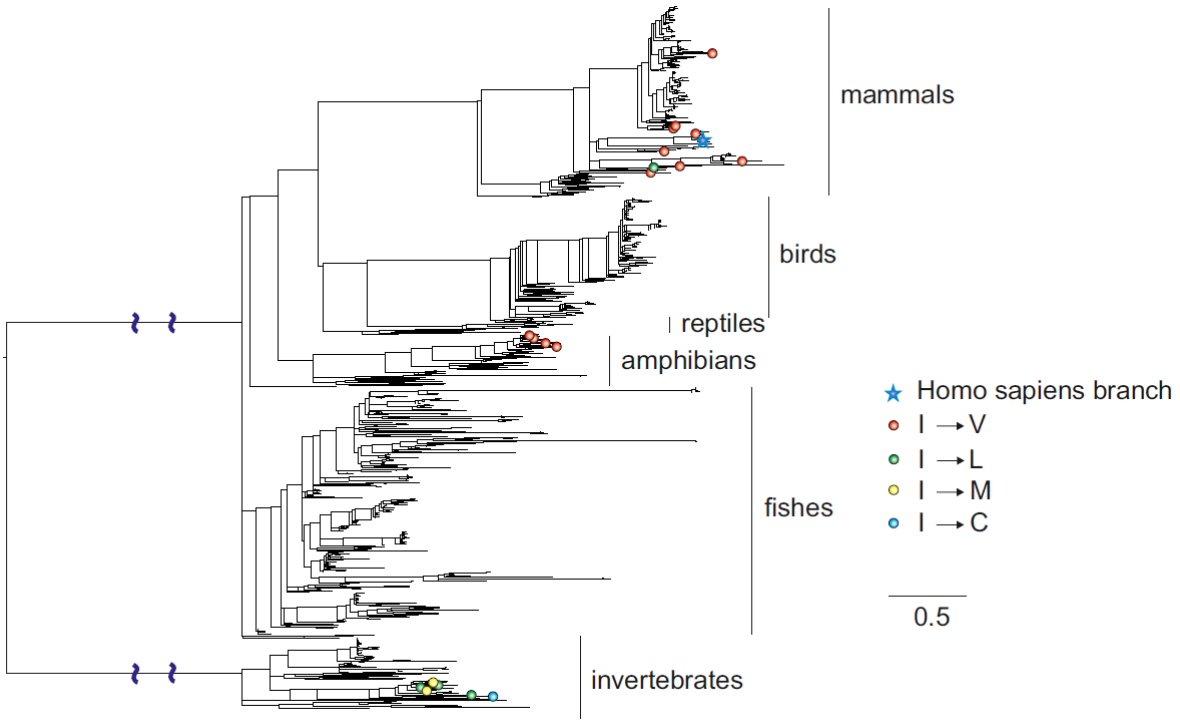



Figure 13