Pathogenic amino acids in mitochondrial proteins more frequently arise in lineages closely related to human than in distant lineages

Galya V Klink [1] , **Georgii A. Bazykin** [Corresp., 1, 2] , **Andrey V. Golovin** [3]

[1] Molecular Evolution, Institute for information transmission problems (Kharkevich Institute), Moscow, Russian Federation
[2] Skolkovo Institute of Science and Technology, Moscow, Russian Federation
[3] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russian Federation

Corresponding Author: Georgii A. Bazykin
Email address: gbazykin@iitp.ru

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. We show that substitutions giving rise to the human amino acid variant 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. Therefore, a pathogenic variant still tends to be more acceptable in human mitochondria than a variant that may only be fit after a substantial perturbation of the protein structure.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3211v1 | CC BY 4.0 Open Access | rec: 31 Aug 2017, publ: 31 Aug 2017

Pathogenic amino acids in mitochondrial proteins more frequently arise in lineages closely related to human than in distant lineages

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

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

[2]Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow 119234, Russia

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

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

Keywords

Fitness landscape, phylogeny, pathogenic mutations, mitochondria, homoplasy

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. We show that substitutions giving rise to the human amino acid variant 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. Therefore, a pathogenic variant still tends to be more acceptable in human mitochondria than a variant that may only be fit after a substantial perturbation of the protein structure.

Significance

Homologous proteins can carry different amino acids at the same position in different species. These changes can be neutral, or can reflect differences in the pressure of selection on these sites. We hypothesized that amino acids observable in human population appear on average more frequently in close than in distant human relatives. For mitochondrial proteins, we observe this for both frequent and rare human alleles. Unexpectedly, substitutions that would be pathogenic in humans also more frequently appear in species more closely related to humans than in distantly related ones. Therefore, despite their pathogenicity, these variants are on average more acceptable in humans than other amino acids that were observed at this site in distantly related species.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3211v1 | CC BY 4.0 Open Access | rec: 31 Aug 2017, publ: 31 Aug 2017

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, Bhaskar, & Pritchard, 2016; Storz, 2016).

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 (Goldstein, Pollard, Shah, & Pollock, 2015; Naumenko, Kondrashov, & Bazykin, 2012; Povolotskaya & Kondrashov, 2010; Rogozin, Thomson, Csürös, Carmel, & Koonin, 2008; Zou & Zhang, 2015).

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 (Jordan et al., 2015; Kondrashov, Sunyaev, & Kondrashov, 2002; Soylemez & Kondrashov, 2012). 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; Klink & Bazykin, 2017; Zou & Zhang, 2015), 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 & Bazykin, 2017) 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.

Materials and Methods

Data

Here, we reanalyzed the phylogenetic data on a set of 5 mitochondrial protein-coding genes in 4350 species of opisthokonts (Klink & Bazykin, 2017), as well as on each of the 12 mitochondrial protein-coding genes in several thousand species of metazoans, focusing on the amino acids present in the human mitochondrial genome.

A joint alignment of five concatenated mitochondrial genes of opisthokonts and alignments of 12 mitochondrial proteins of metazoans (Breen et al. 2012) were obtained as described in (Klink & Bazykin, 2017). These alignments were used to reconstruct constrained phylogenetic trees, ancestral states and phylogenetic positions of substitutions (Klink & Bazykin, 2017). 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

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 & Bazykin, 2017). 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 & Bazykin, 2017).

Molecular dynamics simulation of single point mutations in position 91 of COX3

The amino acid at position 91 of COX3 is adjacent to the pore in the complex IV of the respiratory chain which is thought to be required for the transport of oxygen. Therefore, estimation of the effect of the mutation in this position requires a complete description of the cytochrome c oxidase complex in membrane environment. We applied coarse grain description with martini forcefield (Marrink, Risselada, Yefimov, Tieleman, & de Vries, 2007). Human cytochrome C oxydase was modelled with homology modelling approach using Modeller 9.18 (Eswar et al., 2006). The membrane environment was rebuild from a random position of lipids and restrained protein structure as in MemProt MD database (Stansfeld et al., 2015). Each mutant was subjected to 0.5 mks simulation with two replicas in GROMACS 2016 (Abraham et al., 2015) with time step 10fs with PME electrostatics (Wennberg et al., 2015). Simulation results were analyzed with MDAnalysis Python module (Michaud-Agrawal, Denning, Woolf, & Beckstein, 2011). Water molecules within the pore were counted in a cylindrical selection with radius 2nm, height 5nm and center position determined dynamically as the center of mass of the two helices of the dimer harboring mutations. Counts of water molecules and their standard deviations were estimated from the last 100 ns of the trajectory.

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, using phylogenies reconstructed from two datasets of mitochondrial protein coding genes obtained earlier (Klink & Bazykin, 2017), we asked how homoplasic substitutions giving rise to the human variant are positioned phylogenetically relative to the human branch, compared with other (divergent) substitutions.

The reference human allele is also observed, on average, in over half of other considered species (71.2% of species of opisthokonts for the 5-genes dataset, and 60% of all species of metazoans for the 12-genes dataset). In 7.1% (10%) of these species, it did not share common ancestry with the human, but instead originated independently in an average of 30 (21) independent homoplasic substitutions per site (Table S1 and Fig. S1-S2).

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. 1a; Fig. S3). 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. 1a; Fig. S3). 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. 1b; Fig. S4). 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 1b; Fig. S4).

Figure 1

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 & Bazykin, 2017). 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. S3).

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 (Table S2).

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 2; Fig. S5-S6). 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. 2a; Fig. S5), and the density of homoplasic substitutions giving rise to such an allele dropped significantly with phylogenetic distance from the human branch (Fig 2b; SI, Fig. S6). 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. S5).

Figure 2

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 (Table S3), the variance in the estimates of the H/D ratio is, as expected, large. Still, in the opisthokont dataset (Fig. 3), as well as in five of the twelve genes of the metazoan dataset (Fig. S7-S8), 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. 3b). 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.S7).

Figure 3

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. 4a) and pathogenic (Fig. 4b) human amino acids were more similar in their biochemical properties according to the Miyata matrix (Miyata, Miyazawa, & Yasunaga, 1979) to the human reference variant than amino acids observed in non-human species. In turn, amino acids observed in non-human species were more similar to the human reference variant than amino acids never observed at this site in any species.

Figure 4.

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. 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 (Fig. S9). For the metazoan dataset, we also plot three select individual amino acid sites with known pathogenic mutations which are considered below.

The V❸A 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 & 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. 5). As a result, the mean phylogenetic distance between human and the parallel V❸A substitutions is 2.35 (median 0.75), while it is 4.12 (median 4.6) for substitutions of V to other amino acids.

Figure 5

The V❸A 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 giving rise to other amino acids occurred throughout metazoans (Fig. 6). As a result, the mean phylogenetic distance from human to V❸A 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 6

To better understand the possible reasons for the unexpected pattern of clustering of the deleterious variant in the phylogenetic vicinity of human, we used molecular dynamics simulations to predict the effect of each mutation on the structure and function of the human protein. As site 91 is positioned within the wall of a pore that is thought to be a channel for oxygen transport (Shinzawa-Itoh et al., 2007), we estimated the pore size that would correspond to each amino acid that occured at this site elswhere on the phylogeny if it arose in the mammalian context. All amino acids led to an increase of the pore size, compared to the normal V allele. Such an increase is expected to permit water molecules to enter the pore, and to impede or prevent oxygen transport. Among the eight observed amino acids, the human pathogenic variant A alters the pore size to the smallest extent, while variants observed in other species significantly increase it, potentially interfering with function (Fig. 7).

Figure 7

Finally, the I❸V 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. 8). 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.

Figure 8

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, Hahn, & Hahn, 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 & Bazykin, 2017), 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 populations, suggesting that these mutations are indeed damaging, while they demonstrate a pronounced clustering (Fig. 3b).

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.

Amino acid at position 91 of the COX3 protein provides a case in point. When the mammalian protein structure is used for modelling, we predict that the human pathogenic variant disrupts structure to a smaller extent than each of the seven variants found in reference genomes of other 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, Dudley, Filipski, & Liu, 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.

Conclusions.

At a given amino acid site, an amino acid pathogenic in human can be tolerable in non-human species due to changes in the genomic context – a phenomenon known as compensated pathogenic deviation (Bazykin, 2015; Jordan et al., 2015; Kondrashov et al., 2002). Here, we describe the opposite facet of the same phenomenon: all amino acids observed in humans, including pathogenic ones, are also more likely to arise independently in lineages close to humans, compared to more distantly related species. Therefore, perturbation of the single position fitness landscape in the course of evolution makes the human pathogenic variants even more deleterious, perhaps lethal, in distantly related species.

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.

Acknowledgements

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

References.

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, *1-2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., … Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–249. https://doi.org/10.1038/nmeth0410-248

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., & Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genetics*, *23*(2), 147. https://doi.org/10.1038/13779

Bazykin, G. A. (2015). Changing preferences: deformation of single position amino acid fitness landscapes and evolution of proteins. *Biology Letters*, *11*(10). https://doi.org/10.1098/rsbl.2015.0315

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M., … Sali, A. (2006). Comparative Protein Structure Modeling Using Modeller. In A. Bateman, W. R. Pearson, L. D. Stein, G. D. Stormo, & J. R. Yates (Eds.), *Current Protocols in Bioinformatics* (pp. 5.6.1–5.6.30). Hoboken, NJ, USA: John Wiley & Sons, Inc. Retrieved from http://doi.wiley.com/10.1002/0471250953.bi0506s15

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* (No. biorxiv;030338v2). Retrieved from http://biorxiv.org/lookup/doi/10.1101/030338

Goldstein, R. A., Pollard, S. T., Shah, S. D., & Pollock, D. D. (2015). Nonadaptive Amino Acid Convergence Rates Decrease over Time. *Molecular Biology and Evolution*, *32*(6), 1373–1381. https://doi.org/10.1093/molbev/msv041

Harpak, A., Bhaskar, A., & Pritchard, J. K. (2016). *Effects of variable mutation rates and epistasis on the distribution of allele frequencies in humans* (No. biorxiv;048421v1). Retrieved 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 Research*, *34*(Database issue), D749–751. https://doi.org/10.1093/nar/gkj010

Ji, Y., Liang, M., Zhang, J., Zhang, M., Zhu, J., Meng, X., … Guan, M.-X. (2014). Mitochondrial haplotypes may modulate the phenotypic manifestation of the LHON-associated ND1 G3460A mutation in Chinese families. *Journal of Human Genetics*, *59*(3), 134–140. https://doi.org/10.1038/jhg.2013.134

Jordan, D. M., Frangakis, S. G., Golzio, C., Cassa, C. A., Kurtzberg, J., Task Force for Neonatal Genomics, … Katsanis, N. (2015). Identification of cis-suppression of human disease mutations by comparative genomics. *Nature*, *524*(7564), 225–229. https://doi.org/10.1038/nature14497

Klink, G. V., & Bazykin, G. A. (2017). Parallel Evolution of Metazoan Mitochondrial Proteins. *Genome Biology and Evolution*, *9*(5), 1341–1350. https://doi.org/10.1093/gbe/evx025

Kondrashov, A. S., Sunyaev, S., & Kondrashov, F. A. (2002). Dobzhansky-Muller incompatibilities in protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(23), 14878–14883. https://doi.org/10.1073/pnas.232565499

Kumar, S., Dudley, J. T., Filipski, A., & Liu, L. (2011). Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. *Trends in Genetics: TIG*, *27*(9), 377–386. https://doi.org/10.1016/j.tig.2011.06.004

Lott, M. T., Leipzig, J. N., Derbeneva, O., Xie, H. M., Chalkia, D., Sarmady, M., … Wallace, D. C. (2013). mtDNA Variation and Analysis Using MITOMAP and MITOMASTER. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *1*(123), 1.23.1–1.23.26. https://doi.org/10.1002/0471250953.bi0123s44

Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., & de Vries, A. H. (2007). The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *The Journal of Physical Chemistry B*, *111*(27), 7812–7824. https://doi.org/10.1021/jp071097f

Mendes, F. K., Hahn, Y., & Hahn, M. W. (2016). Gene tree discordance can generate patterns of diminishing convergence over time. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msw197

Michaud-Agrawal, N., Denning, E. J., Woolf, T. B., & Beckstein, O. (2011). MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry*, *32*(10), 2319–2327. https://doi.org/10.1002/jcc.21787

Miyata, T., Miyazawa, S., & Yasunaga, T. (1979). Two types of amino acid substitutions in protein evolution. *Journal of Molecular Evolution*, *12*(3), 219–236.

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. *Journal of Child Neurology*, *26*(1), 12–20. https://doi.org/10.1177/0883073810371227

Munakata, K., Tanaka, M., Mori, K., Washizuka, S., Yoneda, M., Tajima, O., … Kato, T. (2004). Mitochondrial DNA 3644T-->C mutation associated with bipolar disorder. *Genomics*, *84*(6), 1041–1050. https://doi.org/10.1016/j.ygeno.2004.08.015

Naumenko, S. A., Kondrashov, A. S., & Bazykin, G. A. (2012). Fitness conferred by replaced amino acids declines with time. *Biology Letters*, *8*(5), 825–828. https://doi.org/10.1098/rsbl.2012.0356

Povolotskaya, I. S., & Kondrashov, F. A. (2010). Sequence space and the ongoing expansion of the protein universe. *Nature*, *465*(7300), 922–926. https://doi.org/10.1038/nature09105

Poznik, G. D., Henn, B. M., Yee, M.-C., Sliwerska, E., Euskirchen, G. M., Lin, A. A., … Bustamante, C. D. (2013). Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science (New York, N.Y.)*, *341*(6145), 562–565. https://doi.org/10.1126/science.1237619

Rogozin, I. B., Thomson, K., Csürös, M., Carmel, L., & Koonin, E. V. (2008). Homoplasy in genome-wide analysis of rare amino acid replacements: the molecular-evolutionary basis for Vavilov's law of homologous series. *Biology Direct*, *3*, 7. https://doi.org/10.1186/1745-6150-3-7

Shinzawa-Itoh, K., Aoyama, H., Muramoto, K., Terada, H., Kurauchi, T., Tadehara, Y., … Yoshikawa, S. (2007). Structures and physiological roles of 13 integral lipids of bovine heart cytochrome c oxidase. *The EMBO Journal*, *26*(6), 1713–1725. https://doi.org/10.1038/sj.emboj.7601618

Soylemez, O., & Kondrashov, F. A. (2012). Estimating the rate of irreversibility in protein evolution. *Genome Biology and Evolution*, *4*(12), 1213–1222. https://doi.org/10.1093/gbe/evs096

Stansfeld, P. J., Goose, J. E., Caffrey, M., Carpenter, E. P., Parker, J. L., Newstead, S., & Sansom, M. S. P. (2015). MemProtMD: Automated Insertion of Membrane Protein Structures into Explicit Lipid Membranes. *Structure*, *23*(7), 1350–1361. https://doi.org/10.1016/j.str.2015.05.006

Storz, J. F. (2016). Causes of molecular convergence and parallelism in protein evolution. *Nature Reviews. Genetics*, *17*(4), 239–250. https://doi.org/10.1038/nrg.2016.11

Tawata, M., Hayashi, J. I., Isobe, K., Ohkubo, E., Ohtaka, M., Chen, J., … 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*(7), 1269–1272.

Wennberg, C. L., Murtola, T., Páll, S., Abraham, M. J., Hess, B., & Lindahl, E. (2015). Direct-Space Corrections Enable Fast and Accurate Lorentz–Berthelot Combination Rule Lennard-Jones Lattice Summation. *Journal of Chemical Theory and Computation*, *11*(12), 5737–5746. https://doi.org/10.1021/acs.jctc.5b00726

Xie, S., Zhang, J., Sun, J., Zhang, M., Zhao, F., Wei, Q.-P., … Guan, M.-X. (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 A. DNA Mapping, Sequencing, and Analysis*, 1–8. https://doi.org/10.3109/19401736.2015.1136304

Zou, Z., & Zhang, J. (2015). Are Convergent and Parallel Amino Acid Substitutions in Protein Evolution More Prevalent Than Neutral Expectations? *Molecular Biology and Evolution*, *32*(8), 2085–2096. https://doi.org/10.1093/molbev/msv091

Figure 1(on next page)

Phylogenetic distances between the human branch and substitutions to the human reference amino acid.

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 higher fraction of homoplasic substitutions to the considered amino acid, compared with random amino acids that had independently originated at this site (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. a) 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; rand, a random non-human amino acid among those that were present in the site; sim, human allele in simulated data. b) 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.*

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3211v1 | CC BY 4.0 Open Access | rec: 31 Aug 2017, publ: 31 Aug 2017

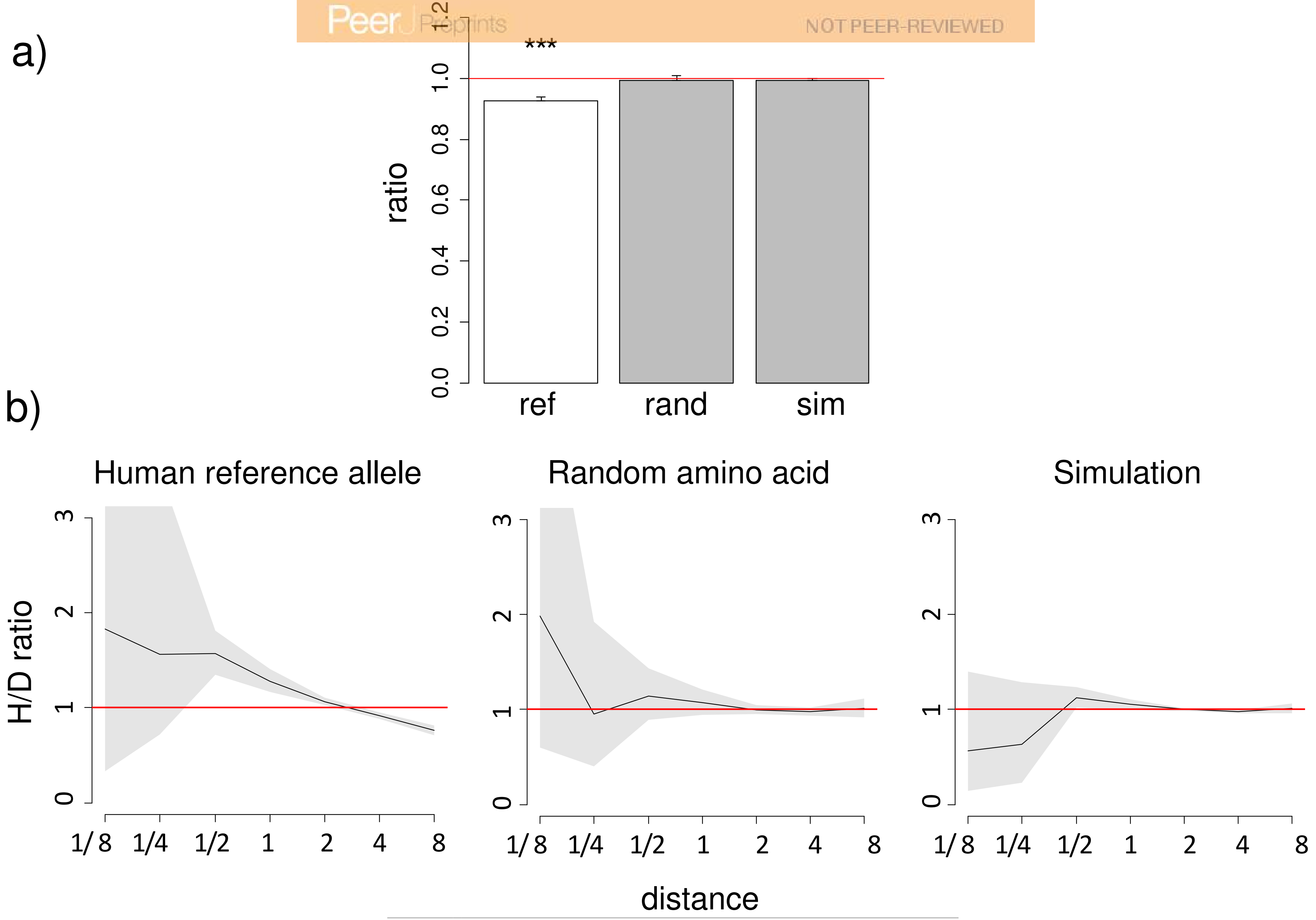

a)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
ratio
***
ref
rand
sim
b)
Human reference allele
Random amino acid
Simulation
H/D ratio
0
1
2
3
1/8
1/4
1/2
1
2
4
8
distance

Figure 2(on next page)

Phylogenetic distances between the human branch and substitutions to the human non-reference amino acid.

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.

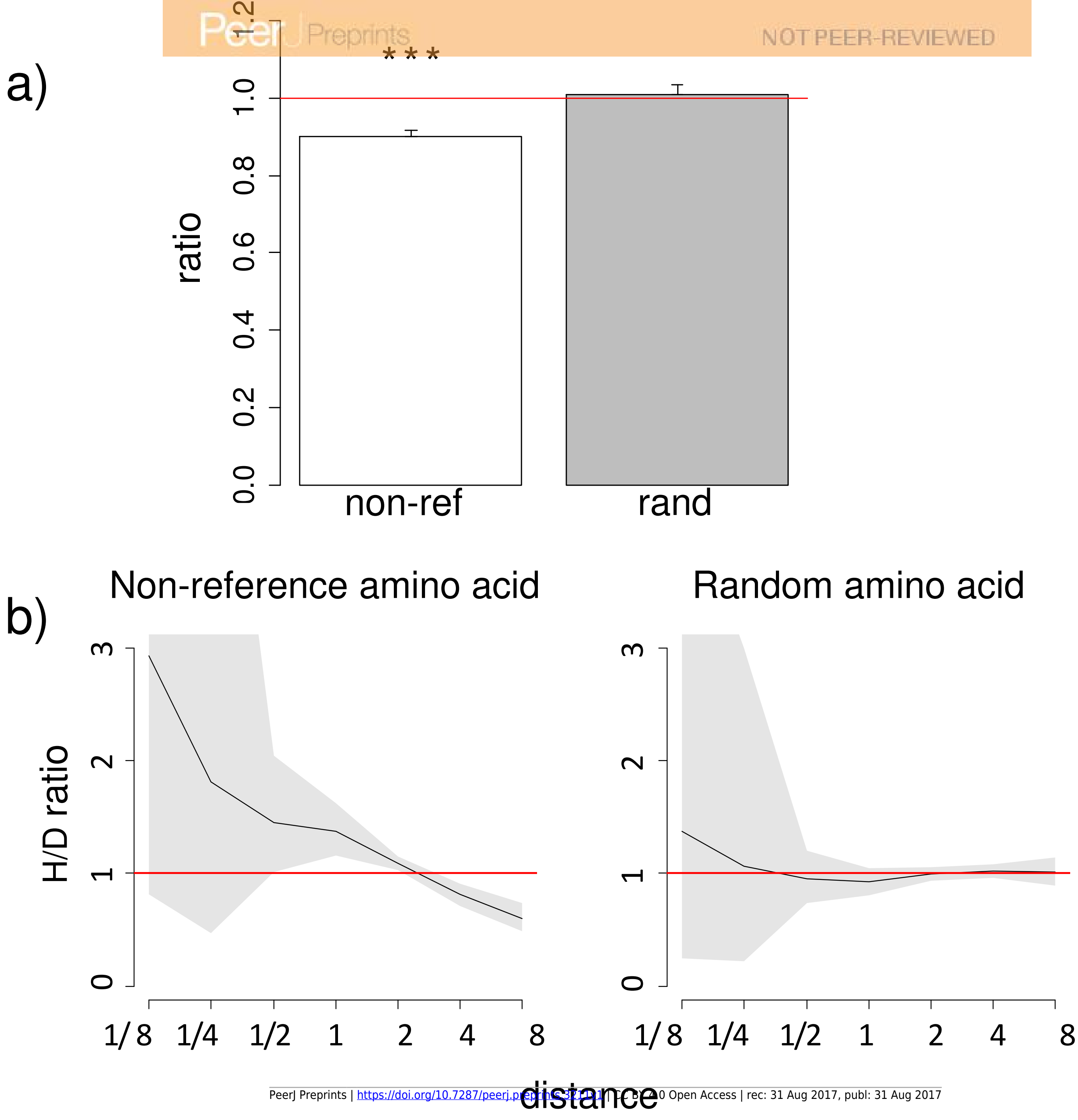

a)
ratio
0.0
0.2
0.4
0.6
0.8
1.0
1.2
* * *
non-ref
rand
b)
Non-reference amino acid
Random amino acid
H/D ratio
0
1
2
3
1/ 8
1/4
1/2
1
2
4
8
distance

Figure 3(on next page)

Phylogenetic distances between the human branch and substitutions to the human pathogenic amino acid.

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.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3211v1 | CC BY 4.0 Open Access | rec: 31 Aug 2017, publ: 31 Aug 2017

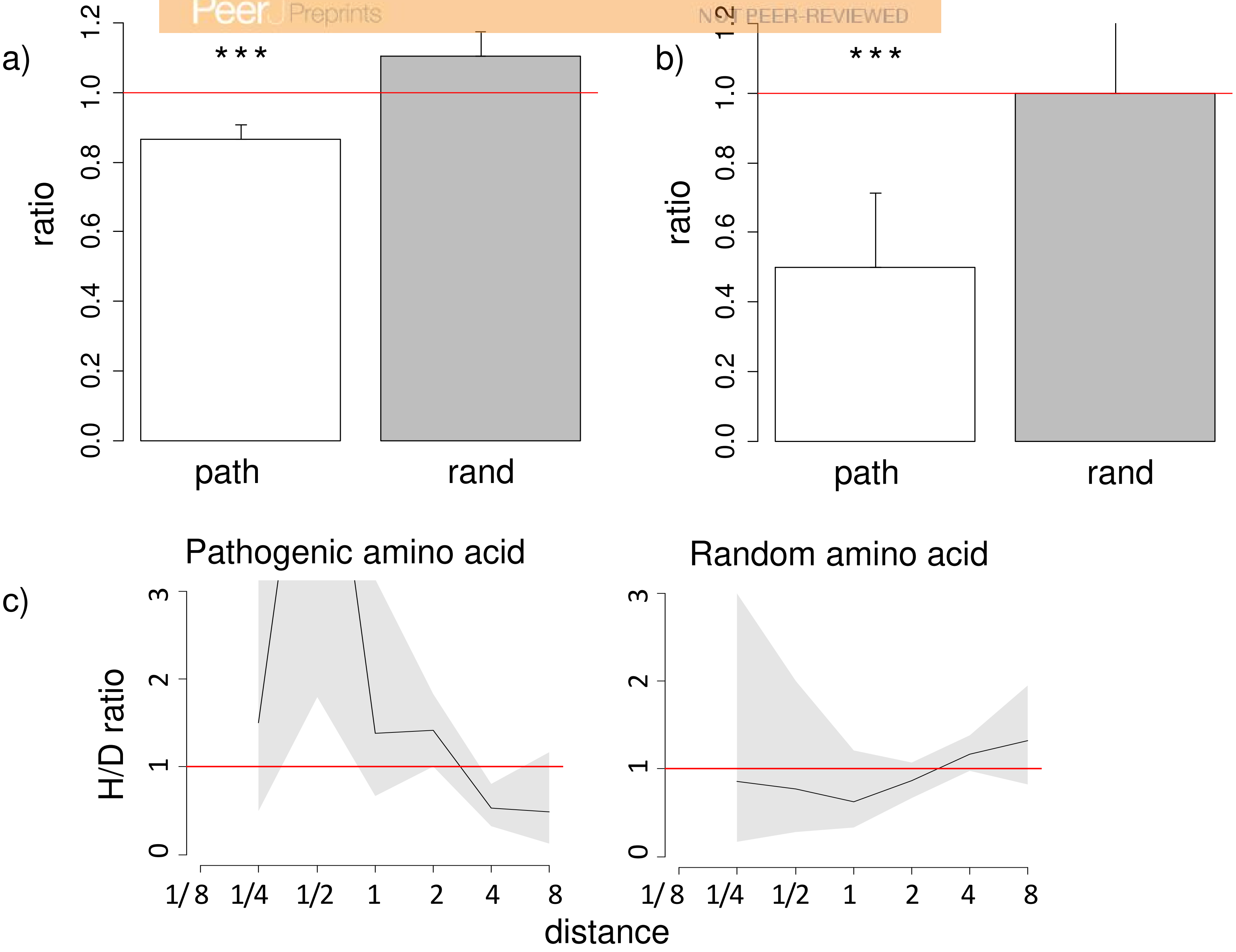
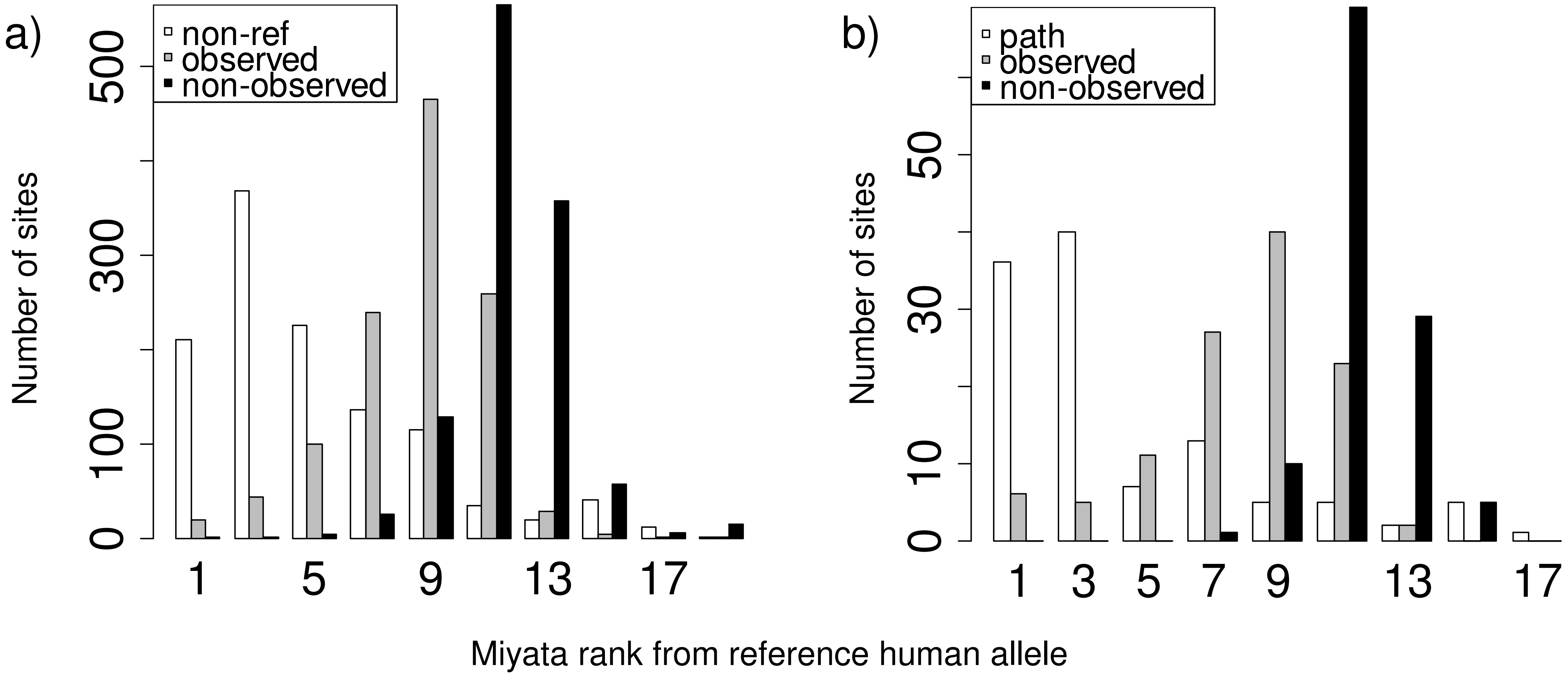

a)
***
ratio
0.0
0.2
0.4
0.6
0.8
1.0
1.2
path
rand
b)
***
ratio
0.0
0.2
0.4
0.6
0.8
1.0
1.2
path
rand
c)
Pathogenic amino acid
H/D ratio
0
1
2
3
1/ 8
1/4
1/2
1
2
4
8
Random amino acid
0
1
2
3
1/ 8
1/4
1/2
1
2
4
8
distance

Figure 4(on next page)

Distributions of ranks of Miyata distances between the reference human allele and the non-reference or pathogenic allele compared with other alleles that were or were not observed at the same site.

White - non-reference (a) or pathogenic (b) allele, gray - other alleles that were observed at the same site, black - other alleles that were not 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.



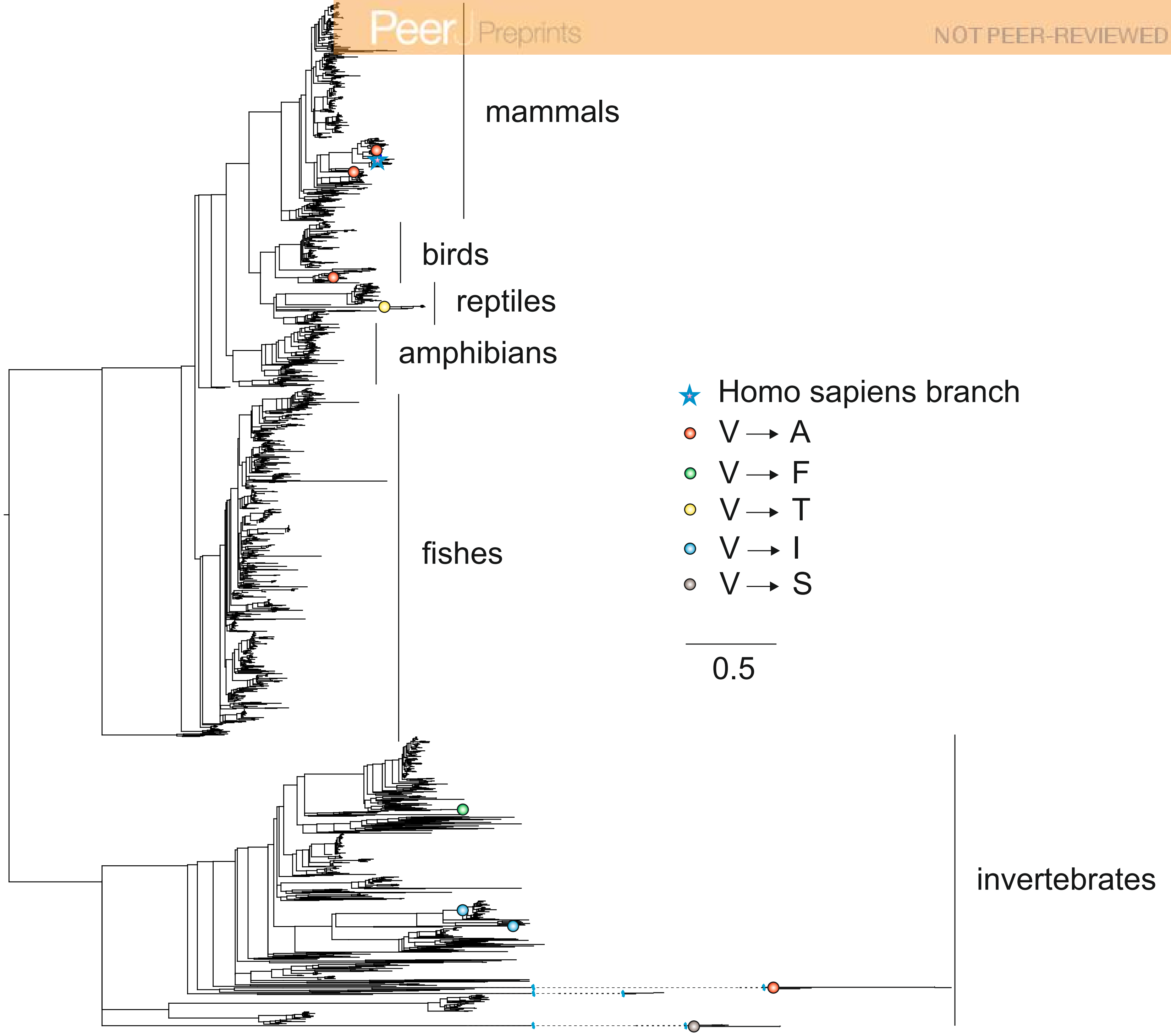
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3211v1 | CC BY 4.0 Open Access | rec: 31 Aug 2017, publ: 31 Aug 2017

Figure 5(on next page)

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.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3211v1 | CC BY 4.0 Open Access | rec: 31 Aug 2017, publ: 31 Aug 2017



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3211v1 | CC BY 4.0 Open Access | rec: 31 Aug 2017, publ: 31 Aug 2017

Figure 6(on next page)

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.

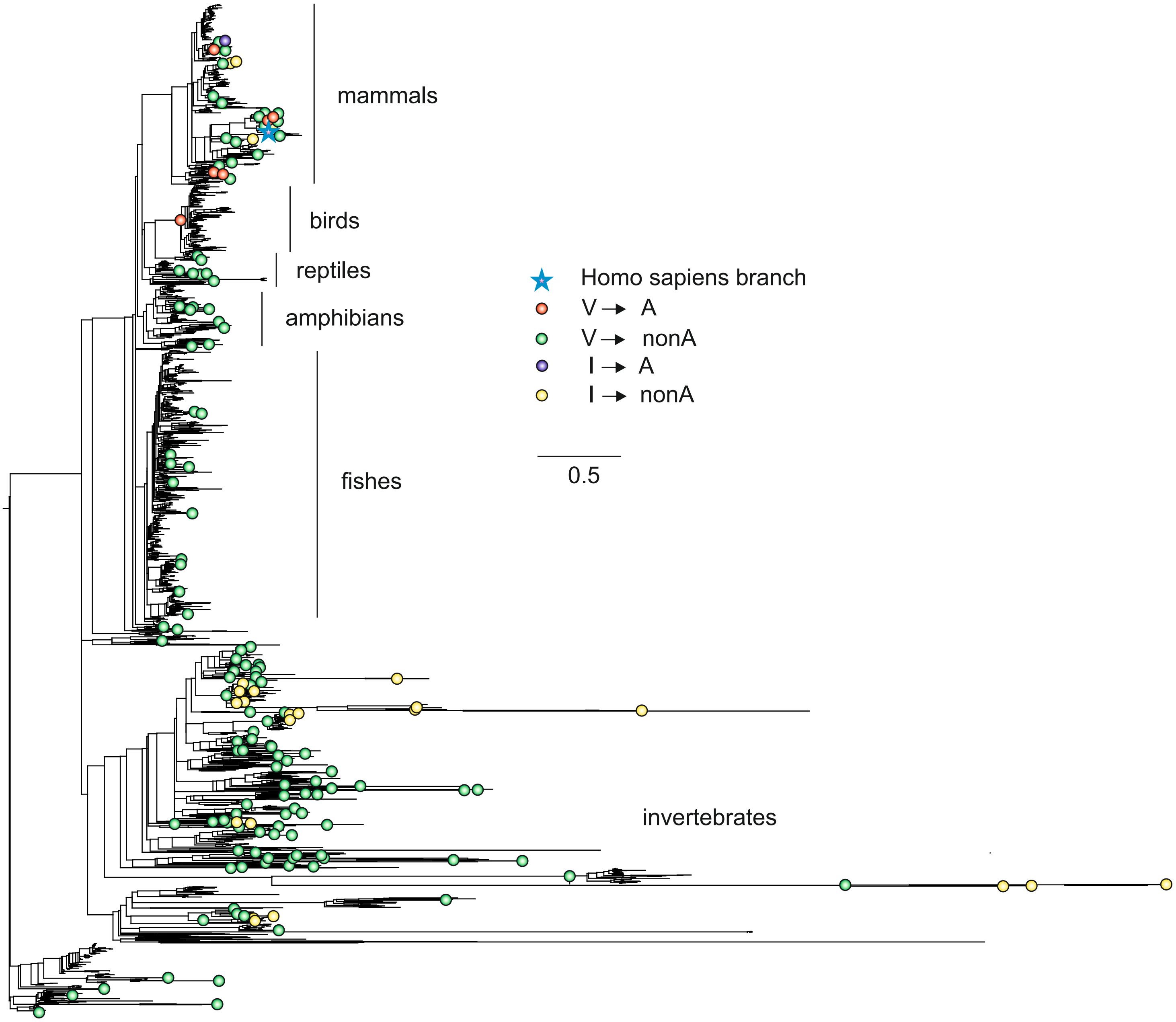

mammals
birds
reptiles
amphibians
fishes
invertebrates
★ Homo sapiens branch
● V → A
● V → nonA
● I → A
● I → nonA
0.5

Figure 7(on next page)

Modelled counts of water molecules in the pore of the human mitochondrial cytochrome c oxidase harboring different mutations in position 91 of COX3.

A higher number of molecules probably impedes or prevents oxygen transport. Blue, human reference amino acid (V); red, human pathogenic amino acid (A); green, amino acids observed in non-human species.

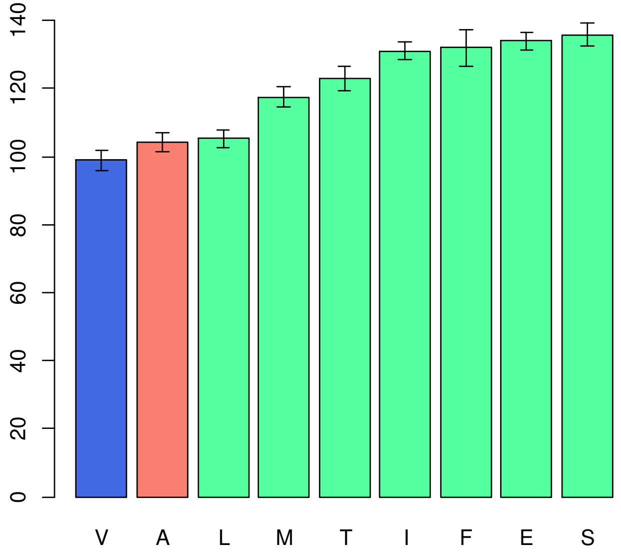
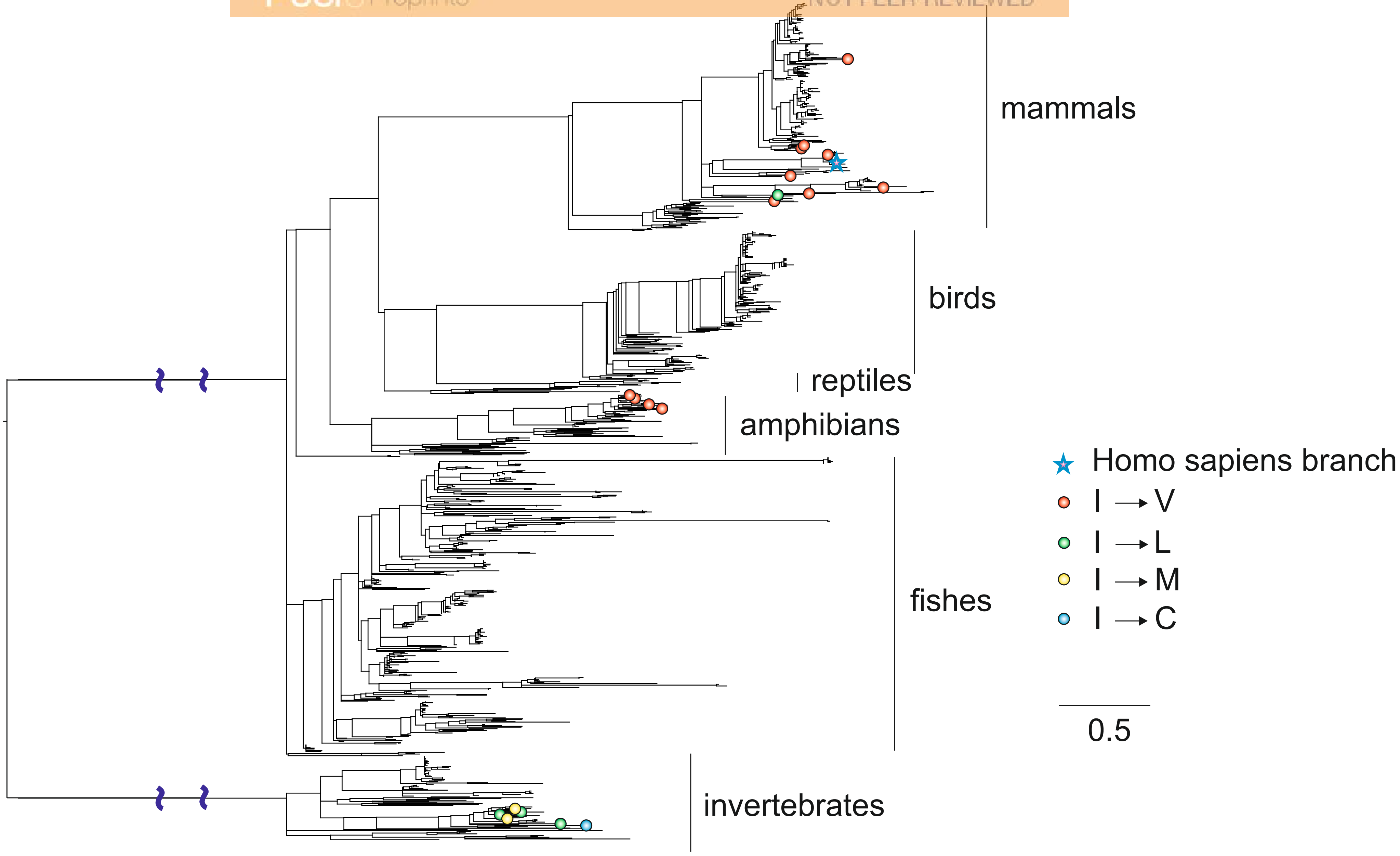

0
20
40
60
80
100
120
140
V
A
L
M
T
I
F
E
S

Figure 8(on next page)

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.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3211v1 | CC BY 4.0 Open Access | rec: 31 Aug 2017, publ: 31 Aug 2017


mammals
birds
reptiles
amphibians
fishes
invertebrates
★ Homo sapiens branch
● I → V
● I → L
● I → M
● I → C
0.5