A novel phylogenetic analysis combined with a machine learning approach predicts human mitochondrial variant pathogenicity

Bala Anı Akpınar[1 †] , Paul O. Carlson[1], and Cory D. Dunn[1 †]

[1]Institute of Biotechnology, Helsinki Institute of Life Science, University of Helsinki, Helsinki, 00014, Finland

[†] Corresponding authors

Correspondence:

Bala Anı Akpınar, Ph.D.
P.O. Box 56
University of Helsinki
00014 Finland
Email: ani.akpinar@helsinki.fi
Phone: +358 50 311 9307

or

Cory Dunn, Ph.D.
P.O. Box 56
University of Helsinki
00014 Finland
Email: cory.dunn@helsinki.fi
Phone: +358 50 311 9307

ABSTRACT

Linking mitochondrial DNA (mtDNA) mutations to patient outcomes has been a serious challenge. The multicopy nature and potential heteroplasmy of the mitochondrial genome, differential distribution of mutant mtDNAs among various tissues, genetic interactions among alleles, and environmental effects can hamper clinicians as they try to inform patients regarding the etiology of their metabolic disease. Multiple sequence alignments using samples ranging across multiple organisms and taxa are often deployed to assess the overall conservation of any site within a mtDNA-encoded macromolecule and to determine the acceptability of any given variant at a particular position. However, the utility of multiple sequence alignments in pathogenicity prediction can be restricted by factors including sample set bias, alignment errors, and sequencing errors. Here, we describe a novel and empirical approach for assessing site-specific conservation and variant acceptability that depends upon phylogenetic analysis and ancestral prediction and minimizes current alignment limitations. Next, we use machine learning to predict the pathogenicity of thousands of so-far-uncharacterized human alleles catalogued in the clinic. Our work demonstrates that a substantial portion of encountered mtDNA alleles not yet characterized as harmful are, in fact, likely to be deleterious. Beyond general applications of our methodology that lie outside of mitochondrial studies, our findings are likely to be of direct relevance to those at risk of mitochondria-associated illness.

INTRODUCTION

Because of the critical role that mitochondria play in metabolism and bioenergetics, mutations of mitochondria-localized proteins and ribonucleic acids can adversely affect human health (Alston *et al.* 2017; Suomalainen and Battersby 2018; Khan *et al.* 2020). Indeed, at least one in 5000 people (Chinnery *et al.* 2000; Gorman *et al.* 2015) is overtly affected by mitochondrial disease. While some mitochondrial DNA (mtDNA) lesions can be clearly linked to mitochondria-associated illness, the clinical outcome for many other mtDNA alleles is more ambiguous (Vento and Pappa 2013). Heteroplasmy among the hundreds of mitochondrial DNA (mtDNA) molecules found within a cell (Stewart and Chinnery 2015; Hahn and Zuryn 2019), differential distribution of disease-causing mtDNA among tissues (Boulet *et al.* 1992), and modifier alleles within the mitochondrial genome (Elliott *et al.* 2008; Wei *et al.* 2017) magnify the problem of interpreting effects of any given mtDNA allele. Mito-nuclear interactions and environmental effects may also determine the outcome of mitochondrial DNA mutations (Wolff *et al.* 2014; Matilainen *et al.* 2017; Hill *et al.* 2019). Beyond the obvious importance of determining the genetic etiology of symptoms as a patient enters the clinic, the rapidly increasing prominence of direct-to-consumer genetic testing (Phillips *et al.* 2018) calls for an improved understanding of which mtDNA polymorphisms might affect human health (Blell and Hunter 2019).

While a continuous increase in the number of human mtDNA sequences provides some indication of which mitochondrial changes may be deleterious, simple tabulation of which mtDNA alleles are or are not found among healthy individuals (Whiffin *et al.* 2017) may, on its own, lack power in predicting mitochondrial disease. Differing, strand-specific mutational propensities for mtDNA nucleotides at different locations within the molecule (Tanaka and Ozawa 1994; Reyes *et al.* 1998; Faith and Pollock 2003) should be taken into account when assessing population-wide data, yet allele frequencies are rarely, if ever, normalized in this way. Moreover, population sampling biases and recent population bottleneck effects can lead to misinterpretation of actual allele frequencies when determining pathogenicity (Keinan and Clark 2012; Zuk *et al.* 2014; Chheda *et al.* 2017; Landry *et al.* 2018). A lack of selection against alleles that might act in a deleterious manner at the post-reproductive stage of life also makes likely the possibility that some mtDNA alleles will contribute to age-related phenotypes while avoiding clear linkage to mitochondrial disease (Medawar 1952; Maklakov *et al.* 2015).

Multiple sequence alignments can offer important assistance when predicting an allele's potential pathogenicity (Raychaudhuri 2011). However, caveats are also associated with predicting allele outcome by the use of these alignments. First, while knowledge of amino acid physio-chemical properties is widely considered to be informative regarding whether an amino acid substitution may or may not have a damaging effect on protein function (Dayhoff *et al.* 1978), the site-specific acceptability of a given substitution is ultimately decided within the context of its local protein environment (Zuckerkandl and Pauling 1965). Second, sampling biases and improper clade selection may deceive the clinician regarding the relative acceptability of a specific allele (Keinan and Clark 2012; Zuk *et al.* 2014; Chheda *et al.* 2017; Landry *et al.* 2018). Third, epistasis between sites, combined with, loosely defined, the 'environment' can further complicate pathogenicity prediction (Kondrashov *et al.* 2002; Starr and Thornton 2016). Finally, alignment errors (Kawrykow *et al.* 2012; Iantorno *et al.* 2014) and sequencing errors (Chen *et al.* 2017; Smith 2019) may falsely indicate acceptability of a particular mtDNA alteration.

Here, we deploy a novel methodology to calculate, with potentially unlimited dynamic range, and in a manner robust against sample bias, sequencing error, and alignment error, the relative conservation of mtDNA-encoded positions. By subsequent application of machine learning, we demonstrate that a surprising number of discovered, yet so-far uncharacterized mtDNA alleles are likely to promote disease.

RESULTS

Mapping apparent substitutions to a phylogenetic tree allows measurement of the relative conservation of positions within a macromolecule

We previously introduced an improved, empirical approach for quantification of substitutions within evolving mtDNA-encoded macromolecules (Dunn *et al.* 2019), and we sought here to develop our methods for prediction of human allele pathogenicity. Toward this goal, we first retrieved full mammalian mtDNA sequences from the National Center for Biotechnology Information Reference Sequence (NCBI RefSeq) Database, then extracted specific sequences of from each RNA or protein coding gene based upon annotation of the *Homo sapiens* reference mtDNA (NC_012920.1). Next, we translated protein-coding sequences using the appropriate codon table, and aligned protein, tRNA, and rRNA sequences (Figure 1a). After concatenating sequences of each species based upon molecule class, we generated maximum likelihood trees (Figure 1b) before performing ancestral prediction (Figure 1c) to reconstruct the character values of each position at each bifurcating node within each tree. Subsequently, using the sequences of extant species and the predicted ancestral node values, we analysed each edge of the tree for the presence or absence of a mutation at each aligned human position (Figure 1d), since using character frequencies as a proxy of conservation is sensitive to sample biases among extant sequences. Subsequently, we centered our attention upon mutations localized between internal nodes of maximum likelihood trees (Figure 1e), excluding edges leading directly to extant sequences and thereby minimizing the effects of alignment errors and sequencing errors that may lead to misinterpretations of variant pathogenicity. Moreover, mutations mapped to internal edges are more likely to represent fixed changes informative for the purposes of disease prediction, while polymorphisms that have not yet been subject to selection of sufficient strength or duration might be expected to lack predictive power. Inferred ancestral and descendent characters at each edge were extracted following generation of the substitution mapping matrix (Figure 1f). While our approach was first applied to mammalian sequence data, we subsequently performed the same analysis on a more limited set of primate sequences.

Emerging from this workflow, we introduce a metric, SumSub(Int), for which we summed the number of these positional substitutions localized to internal edges of the tree. The SumSub(Int) value provides a readout of relative mutational propensity and, consequently, conservation. To ensure useful predictions regarding the relative conservation of each position with regard to substitution, we limited our analysis to alignment columns with infrequent gaps. When plotting increasing SumSub(Int) values across all amino acid positions with gap values of 1% or less, we noted a characteristic curve for both mammals (Figure 2a) and primates (Figure 2b). We first noted many positions that were nearly or completely invariant, implying critical roles for the folding or function of each polypeptide. Next, there was a slope of increasing SumSub(Int) values associated with amino acids of decreasing evolutionary constraint. At the end of each curve, there is a drastic increase in the slope associated with degenerate positions apparently under minimal or no selection. Similar curves are apparent when examining SumSub(Int) distributions associated with all rRNA- and tRNA-encoding nucleotides (Figures 2c - 2f) and when studying individual macromolecules within each class (Figures S1 - S3).

Kolmogorov-Smirnov analysis of SumSub(Int) distributions for mammalian mtDNA-encoded proteins confirmed (da Fonseca *et al.* 2008) that there are often differences among mtDNA-encoded subunits when considering the number of positions under evolutionary constraint and/or the strength of selection upon those positions (Figures 3a and 3b). Importantly, our novel methodology supports previous data (da Fonseca *et al.* 2008; Nabholz *et al.* 2013) suggesting that the core subunits of cytochrome *c* oxidase (COX), and particularly the COI polypeptide, are the most highly conserved when considering mutability across amino acid positions (Figure 3c). We note the possibility that plots of SumSub(Int) values, and the statistical comparison of the distributions between two clades, may be an excellent methodology for detection of relaxed selection (B.A. Akpinar and C.D. Dunn, manuscript in preparation).

To further demonstrate that low SumSub(Int) values correspond with conservation, we plotted mammalian SumSub(Int) sums for the three mitochondria-encoded subunits of cytochrome c oxidase (COX) onto the high-resolution structure of the bovine enzyme (Yoshikawa *et al.* 1998). Indeed, according to SumSub(Int) values extracted from the mammalian mtDNA tree, conserved positions were localized to the catalytic core of COX (Figure 3d and Movie S1). In contrast, more divergent amino acids faced toward external regions of the complex, especially in regions contacting the lipids of the mitochondrial inner membrane.

Further advancing the idea that SumSub(Int) values are of utility in comparing conservation across sites of a macromolecule, we considered whether turning away from model- or simulation-based substitution rates and toward empirical measurement of mutations, as we have done here, would lead to the undercounting of substitutions if the evolutionary divergence between two nodes in too large. To test the extent to which our method may undercount substitutions at aligned sites, we first selected alignment positions that were equal to or less than 1% gapped, then selected those sites of protein, rRNA, or tRNA that were under the most minimal selective constraint, ranking in the top 1% of SumSub(Int) values. We then compared the number of apparent changes summed across all positions along an internal tree edge to the number of substitutions occurring only in those selected, degenerate sites. The binary nature of our edge-focused substitution analysis, like other approaches investigating substitution, obscures the development of any Poisson distribution (Zuckerkandl and Pauling 1965; Goldman 1990) based on repeated degenerate site substitutions, so we expected the total number of degenerate sites mutated along an edge to plateau with increased evolutionary distance. Indeed, when an edge contained ~150 - 200 mtDNA-encoded protein positions with substitutions, the relationship between total degenerate sites substituted and the total substitutions along the edge changed from linear to asymptotic when examining mammalian (Figure 4a) or primate (Figure S4a) protein data. The relationship between tRNA and rRNA degenerate site substitutions and total edge substitutions is less clear-cut than for proteins (Figure S4b-S4e), potentially as a result of a lower number of degenerate positions available for analysis and of fewer substitutions at selected degenerate sites. Because the vast majority of edges (~90%) of our analyzed mammal and primate internal tree edges are characterized by less than 150 protein changes, and because most polypeptide positions encoded by mtDNA appear to be evolutionarily constrained and cannot reach the substitution levels characterizing the most degenerate sites (Figure 2), we expect that any potential under-counting of substitutions along longer edges will minimally affect our analysis of site-specific conservation within mtDNA.

Summing apparent substitutions across a phylogenetic tree can provide a measurement of evolutionary divergence

While SumSub(Int) values clearly provide a measure of relative selection when comparing different positions within the same type of mtDNA-encoded macromolecule, results obtained above also indicate that maximal SumSub(Int) values can scale with evolutionary divergence when moving from one taxonomic rank to another. When considering the set of aligned positions that are less than 1% gapped and are ranked within the top 2% of SumSub(Int) values, the median value was approximately six-fold higher in mammals than primates for protein coding positions and nearly five-fold higher for both rRNA and tRNA positions (Figure 4b). While measures of sequence divergence certainly need not correspond with chronological time (Kumar 2005), these values are consistent with the emergence of mammals roughly 200 million years ago (Upham *et al.* 2019) and the divergence of primates from other mammals approximately 70 million years ago (Pozzi *et al.* 2014; Reis *et al.* 2018). While our primary purpose here will be to predict the pathogenicity of mtDNA mutations, SumSub(Int) values from well-aligned, yet degenerate sites along edges of a phylogenetic tree should be an important supplement to existing molecular clock methods. In this case, useful comparison of evolutionary divergence using our approach would require sufficient mtDNA sampling across clades of interest.

Conservation values obtained from substitution sums are not limited in dynamic range

Measures of conservation based upon character frequencies extracted from multiple sequence alignments, such as Shannon entropy, can be limited in dynamic range. Limits on dynamic range that can characterize frequency-based approaches to position analysis in alignments are especially prominent when considering nucleotides, where only four possibilities exist at any given position. In contrast, given the demonstrated scaling of SumSub(Int) values with taxonomic rank, summing positional substitutions across edges of a phylogenetic tree should circumvent limitations on the dynamic range of calculated conservation values. To demonstrate the increased dynamic range provided by our method, we compared Shannon entropy values and SumSub(Int) values for well-aligned protein (Figure 5a), rRNA (Figure 5b), and tRNA (Figure 5c) positions encoded by mammal and primate mtDNA. Since Shannon entropy values and SumSub(Int) values can both report upon conservation at aligned positions, these two variables were, as expected, correlated with one another. However, while the range of Shannon entropy values cannot increase when moving from primate-level analysis to mammal-level analysis, the range of SumSub(Int) values did increase, suggesting that our approach to comparing selection at different sites will benefit from the continuous accumulation of mtDNA sequence information.

Low substitution at specific positions along phylogenetic tree edges is linked to human mtDNA allele pathogenicity

Since summation of detected substitutions across a phylogenetic tree provided a robust measure of relative conservation at different macromolecular positions, we were confident that a phylogenetic analysis that includes SumSub(Int) values would be informative about the pathogenicity of human mtDNA alleles. To test this possibility, we focused our attention upon specific variants of protein and RNA encoding genes annotated within the MITOMAP database of human mtDNA alleles (Lott *et al.* 2013). Indeed, we detected a clear relationship between the confirmed pathogenicity of mutations at mitochondria-encoded proteins and the conservation at amino acid positions, as reflected by SumSub(Int) values extracted from mammal (Figure 6a) and primate (Figure 6b) trees. Similarly, there was a strong link between tRNA mutation pathogenicity and SumSub(Int) values obtained from mammal (Figure 6c) and primate (Figure 6d) trees. The paucity of confirmed pathogenic mitochondrial rRNA mutations in the MITOMAP database made comparisons using this class of molecules impractical, yet future confirmation of additional pathogenic mutations in mitochondria-encoded rRNAs is expected to permit future analyses. Together, findings obtained by phylogenetic analysis of mitochondria-encoded proteins and tRNAs indicate that SumSub(Int) values will indeed be of valuable assistance in the prediction of which mtDNA mutations may lead to disease.

Many mtDNA polymorphisms lead to uncharacterized pathogenicity

We noted that the alignment positions of a subset of alleles currently annotated as ‘polymorphic’ in the MITOMAP database were distinguished by very low SumSub(Int) scores, suggesting that mutations at those positions would, in fact, be deleterious. In order to examine whether additional pathogenic mutations may exist among so-called ‘polymorphic’ alleles in MITOMAP, we devised an approach reliant upon our edge-wise mapping of apparent substitutions, yet only indirectly related to our SumSub(Int) values. Specifically, we looked to the predicted ancestral and descendent characters at mutations mapped to internal tree edges. If direct substitution from the human reference character to the mutant character (or the inverse, assuming the time-reversibility of evolution) is predicted internal to a tree, than such a change at a given position might be expected to be less deleterious than a substitution to or from the human character that was never encountered during evolution. Measuring selection by assessment of population allele frequencies is problematic due to potential sampling biases and population bottlenecks (Auer and Lettre 2015), timing of mutation arrival within an expanding population (Luria and Delbrück 1943), and the divergent nucleotide- and strand-specific mutational propensities of mtDNA (Tanaka and Ozawa 1994; Reyes *et al.* 1998; Faith and Pollock 2003). Even so, we tested whether those MITOMAP-extracted protein mutations which cannot be directly linked to the human character in the mammal phylogenetic tree would be deleterious and might lead, by purifying selection, to a consequent reduction in the number GenBank full-length mtDNA samples harboring the mutation. Indeed, the distribution of MITOMAP samples was significantly different for those mutations for which a direct substitution could be identified in the mammalian tree than for those for which such a substitution could not be identified (Figure 7), strongly supporting the idea that many alleles currently considered polymorphic are harmful and may cause illness.

Graphs of amino acids possible at each tested protein position (regardless of mtDNA sequence gaps) that emerge from analysis of direct substitutions within mtDNA-encoded Complex I (Figure S5) or Complex III, IV, and V (Figure S6), as predicted by use of mammalian data, are provided. Apparent direct substitutions to or from the human reference amino acid in mammals and primates derived from analyzing internal tree edges (Tables S1 and S2) or all tree edges (Tables S3 and S4) are also provided in tabular form. Moreover, we provide tables indicating the simple presence or absence of each amino acid at aligned human positions, as derived from internal tree nodes (including the predicted root) (Tables S5 and S6) or all nodes (Tables S7 and S8).

Deployment of a support vector machine to predict deleterious mtDNA polymorphisms

After finding that SumSub(Int) scores could help predict pathogenicity, and given the clear presence of harmful alleles among uncharacterized polymorphisms, we sought a high-throughput method that would identify potentially pathogenic polymorphisms. We turned to machine learning, and specifically a support vector machine (SVM) (Cortes and Vapnik 1995), to predict the risk of mtDNA polymorphisms detected in full-length human mtDNA sequences. Along with the mutations enumerated in MITOMAP, we also included mtDNA changes found in the recently established HelixMT database (HelixMTdb), which harbors the results of a new genetic survey, performed agnostic to phenotype, of nearly 200,000 people (Bolze *et al.* 2019). Positive training sets for protein consisted of mutations listed as confirmed pathogenic alterations within the MITOMAP database. Positive training sets for tRNA included confirmed deleterious alleles, as well as several alleles reported, but not confirmed, to be pathogenic. Negative training sets included those variants both annotated as polymorphic in MITOMAP and characterized by the highest counts of homoplasmic alleles in HelixMTdb, interpreted here as a sign of benignancy. Analyzed positions were gapped at 1.5% of sequences or less. We did not limit our input SVM features to SumSub(Int) scores, but expanded our features to SumSub(All), which includes edges leading to extant mtDNA sequences; the number of characters found at alignment positions (internal edges or all edges); direct substitution between the human reference character and the mutant human character within analyzed trees (internal edges or all edges); whether the mutant character is found at all in internal nodes or all nodes; and positional Shannon entropy.

For alleles found within protein-coding genes, a SVM clearly permitted very reliable and automated predictions regarding which polymorphisms not yet classified as pathogenic may cause or contribute to disease. When analyzing data from mammalian (Figures 8a and 8b) or primate (Figures S7a and S7b) phylogenies, few confirmed pathogenic mutations were found among the predicted negative set. Importantly, when examining the behavior of alleles in our negative training sets (Figures 8c and S7c), no false-positives were predicted, suggesting that those test alleles predicted as pathogenic will have a high likelihood of causing effects in humans. Error rates for prediction dependent upon mammalian or primate datasets averaged less than 5%. No specific feature stood out as dominant in providing predictive value (Figure 8d and S7d), although this might be expected due to the lack of true independence of any given feature from the others (all, to some degree, are affected by site conservation). The robustness of our model in correctly identifying the true positives in the training set without corresponding loss in specificity (or false positive rate) is highlighted in the Receiver Operating Characteristic (ROC) curves of mammalian (Figure 8e) and primate (Figure S7e) protein predictions (Figure S7e).

For tRNAs, a SVM-based approach provided predictive value, yet analysis of data rooted in mammal (Figure S8) or primate (Figure S9) datasets did not provide the superlative separation of negative and positive training sets that characterized the protein-coding alleles. ROC curves also suggest more modest predictive power of our SVM in predicting tRNA allele pathogenicity. The lack of confirmed pathogenic mutations in rRNA annotated in MITOMAP prevented predictive approaches. Protein (Tables S9 and S10) and tRNA (Tables S11 and S12) pathogenicity predictions are also provided in tabular form.

While our positive training set consisted of alleles of confirmed pathogenicity, there are numerous other alleles provided in the MITOMAP database that are reported, yet not confirmed, to be pathogenic. Protein-coding alleles of this type were distributed on both sides of the SVM decision boundary (Figure 8a and S7a). Our machine learning approach predicted 38% (n = 227) of reported disease-associated alleles to be pathogenic when using the mammalian phylogenetic dataset (Figure 8f), and 42% (n = 213) of reported disease-associated alleles were predicted to be pathogenic when using the primate phylogenetic dataset (Figure S7f). We note that MITOMAP annotations of these alleles of reported pathogenicity encompass clinical outcomes ranging from neuromuscular syndromes to consequences, like prostate cancer, not typically associated with mitochondrial disease. Given the very low false positive rate associated with our approach to predicting pathogenicity of alleles in protein-coding genes, our SVM approach provides strong additional support for nearly 100 alleles already suspected to cause disease.

Since our training sets were initially gleaned from the MITOMAP database, we next turned to alleles reported only in HelixMTdb for further validation of our SVM prediction output. First, we expected that pathogenic alleles are more likely to be detected as heteroplasmic within human samples, since overt mitochondrial disease can require a deleterious allele to rise to a high proportion of the mtDNA molecules maintained by the cell (Stewart and Chinnery 2015; Hahn and Zuryn 2019). Strikingly, we found that for 69% of alleles that we predicted to be pathogenic using mammalian phylogenetic data (n = 640) and for 68% of alleles that we predicted to be pathogenic using primate phylogenetic data (n = 706), all samples harvested were heteroplasmic (Figure 9a). In contrast, for samples predicted as non-pathogenic, the number of alleles characterized by 100% homoplasmy was nearly six-fold lower: 12% for samples analyzed using mammal (n = 5877) or primate (n = 5575) phylogenetic data. The distribution of the fraction of samples heteroplasmic for each allele also differed significantly for those alleles that we predicted as pathogenic versus those predicted as not pathogenic (Figures 9b and 9c). Finally, the number of total samples in HelixMTdb harboring any given allele differed substantially between those predicted as pathogenic and those predicted as non-pathogenic (Figures 9d and 9e), suggesting purifying selection of now predicted, but not yet reported or confirmed, pathogenic mutations within the human population. Together, these analyses strongly support the utility of our SVM predictions in determining the pathogenicity of alleles affecting mtDNA-encoded polypeptides.

DISCUSSION

Establishment of a novel and empirical approach to quantifying conservation and evolutionary divergence

We describe here the elaboration of a new methodology allowing quantification of the relative conservation of sites within and between genes, RNAs, and proteins. Rooted in the use of existing methods of phylogenetic tree generation and ancestral prediction, mapping subtitutions to phylogenetic tree edges can minimize errors resulting from sampling biases, from sequencing errors, and from alignment errors when calculating conservation. Moreover, our approach provides a measure of conservation that is theoretically unlimited in dynamic range and, consequently, can continue to benefit from the ever-increasing mtDNA sequence information available in public databases. Even nearly identical sequences can be utilized by our approach, allowing for a larger input dataset that can be deployed toward calculation of site-specific conservation. Moreover, SumSub(Int) values at degenerate sites can provide a measure of evolutionary divergence that should compliment other approaches dedicated to the calculation of branch length.

Researchers often refer to generalized substitution matrices when predicting whether a change may be harmful or not (Dayhoff *et al.* 1978; Jones *et al.* 1992), yet amino acid exchangeability matrices change across clades (Zou and Zhang 2019), and successful substitution of any given character obviously occurs in the context of a very specific local environment (Zuckerkandl and Pauling 1965). By mapping substitutions to phylogenetic tree edges, ample sequence data can allow determination of character acceptability within the context of a particular macromolecular position, thereby improving prediction of variant pathogenicity. Accordingly, we have generated look-up tables emerging from the prediction of direct substitutions at well-supported internal edges within the mammalian and primate phylogenetic trees. These data will be, beyond application in our automated approaches to pathogenicity prediction, of great use to clinicians and genetic counselors when determining the potential outcome of detected mtDNA alleles.

While we have currently taken advantage of the relatively abundant mtDNA sequence data available, the continuous compilation of nuclear sequences obtained by advancing next-generation sequencing technologies will also allow study of nucleus-encoded genes relevant to disease. We note that our approach is, of course, not limited to the study of mitochondria-encoded components, but can be applied to any macromolecule for which sufficient sequence exists toward generation of a high quality alignment, a well-supported phylogenetic tree, and reliable ancestral predictions.

Machine learning predicts polymorphism pathogenicity

Protein and tRNA alleles of confirmed pathogenicity were clearly characterized by low SumSub(Int) values. Furthermore, a lack of apparent direct substitutions mimicking protein-coding mutations from the human reference sequence could also be indirectly linked to pathogenicity. But how should one weigh these and other available factors when classifying, in a high-throughput manner, the many polymorphisms already encountered or waiting to be discovered during the sequencing human mtDNAs? We deployed a machine learning approach, trained on existing data in the MITOMAP database, toward determination of mtDNA mutation pathogenicity. For alleles found in protein-coding genes, our methodology appeared effective, in that predicted deleterious alleles obtained from a second, independent database of mtDNA polymorphisms showed strong signs of indeed being pathogenic. Specifically, our false-positive rate for protein-coding alleles appeared to be practically non-existent, predicted pathogenic protein mutations were characterized by low prominence in the population, and there was a strong bias against homoplasmy when considering polypeptide mutations that we predicted to be harmful. SVM prediction efficiency for tRNA alleles, based upon receiver operating characteristic curves, appeared to be more moderate. Confirmed disease-causing alleles appear to be disproportionately localized to mitochondrial tRNA genes (Florentz *et al.* 2003; Tuppen *et al.* 2010). Consequently, it is possible that a greater fraction of mutations in analyzed positions are deleterious than for protein-coding sequence, even when the number of clinical samples harboring the allele is high, thereby leading to incorporation of false negatives within our negative training sets. Moreover, while some tRNAs appear to more readily accept nucleotide substitutions, mutations upon other tRNAs appear to be more readily associated with disease (Florentz *et al.* 2003; Wittenhagen and Kelley 2003), suggesting further sub-classification of tRNA molecules may improve pathogenicity prediction. Many tRNA mutations occuring at second sites within the molecule can act in a compensatory manner (Kern and Kondrashov 2004), potentially suppressing pathogenicity of otherwise harmful alleles, and therefore leading to false positives during prediction. Prediction of rRNA allele pathogenicity was not attempted due to the scarcity of confirmed, disease-associated rRNA alleles encoded within human mitochondria. Automated predictions focused upon both classes of RNA molecules should greatly improve upon the development of improved training sets, and our phylogeny-based metrics will serve as valuable resources for clinicians and geneticists in predicting pathogenicity.

Previously uncharacterized mtDNA alleles may lead to cryptic and age-related mitochondrial disease

Our examination of MITOMAP and HelixMTdb entries strongly suggests that a substantial number of deleterious human mutations remain to be classified as pathogenic, so why have these potentially harmful alleles not yet been classified as harmful within the clinic?

First, if a mutation is too common, clinicians may quickly consider the allele to be unlinked to patient symptoms (Whiffin *et al.* 2017). Our findings do, in fact, support the idea that allele population frequency can correspond with pathogenicity. However, population frequency cannot be a dependable predictor of clinical outcome, since, as mentioned above, population counts need to be heavily corrected for strand- and nucleotide-specific mtDNA mutational biases (Tanaka and Ozawa 1994; Reyes *et al.* 1998; Faith and Pollock 2003), and sampling biases are typically a hazard when carrying out population-wide studies (Keinan and Clark 2012; Zuk *et al.* 2014; Chheda *et al.* 2017; Landry *et al.* 2018).

Second, deleterious mtDNA alleles often must rise beyond a certain threshold among the hundreds of mtDNA molecules potentially resident within a cell in order for overt symptoms to manifest. Concordantly, our data suggest a strong propensity for heteroplasmy in the set of alleles that we predict to be pathogenic, but are not yet clinically annotated. Differential distribution of these alleles, either during development or other bottleneck effects in both non-mitotic and renewable tissues (Stewart and Chinnery 2015; Zhang *et al.* 2018), may generate clones with a high proportion of deleterious mutations and to complex, tissue-specific outcomes (Nekhaeva *et al.* 2002; Fayet *et al.* 2002; Greaves *et al.* 2006; Bratic and Larsson 2013). Moreover, the phenomena described above may lead to age-related symptoms not easily classified as mitochondrial disease, since even relatively common mtDNA sequence variants have been suggested to contribute to diseases like diabetes, heart disease, and cancer (Marom *et al.* 2017; Wei *et al.* 2017; Chinnery and Gomez-Duran 2018). We are certainly tantalized by the prospect that pathogenic alleles illuminated by our approach might impinge upon human lifespan.

METHODOLOGY

Mitochondrial DNA sequence acquisition and conservation analysis

Mammalian mtDNA sequences were retrieved from the National Center for Biotechnology Information database of organelle genomes (https://www.ncbi.nlm.nih.gov/genome/browse#!/organelles/ on September 26, 2019). These 1184 mammalian mtDNA genomes were aligned using MAFFT on the ‘auto’ setting (Katoh and Standley 2013). Four sequences that were egregiously misaligned were removed, and MAFFT alignment with the ‘auto’ setting was carried out again. Individual gene sequences were extracted from these alignments, based upon the annotated human mtDNA (NC_012920.1). After gap removal, translation of protein coding genes was performed using the vertebrate mitochondrial codon table in AliView (Larsson 2014). MAFFT alignments of each gene product were performed using the G-INS-i iterative refinement method, then individual concatenates for each species were generated from protein coding sequences, tRNAs, and rRNAs. Duplicates were removed from the protein, tRNA, and rRNA concatenates using seqkit v0.10.2 (Shen *et al.* 2016). Primate subsets containing concatenates from each molecule class were generated after analysis of species names using the taxize package (v0.9.9) in R (Chamberlain and Szöcs 2013).

Maximum likelihood trees for each molecule class concatenate were built using FastTreeMP (Price *et al.* 2010) with four subtree-prune-regraft moves, two rounds of branch length optimization, slow nearest neighbor interchange, and a generalized time-reversible model. Next, ancestral prediction was carried out using the PAGAN package (Löytynoja *et al.* 2012), with concatenated alignments and phylogenetic trees used as input. The PAGAN output files for tRNAs and rRNAs were, before inferring a substitution, processed so that each ambiguous IUPAC nucleotide code was replaced by one of the possible standard nucleotides represented by that IUPAC code within the entire tree at that given position. This process was carried out using using script “replace-nonstandard.py” (https://github.com/corydunnlab/Edge_mapping_conservation), where the standard nucleotide was picked at random from available standard characters at a given alignment position using the random.choice() function of the Python random module. For positions that contained only N’s and gaps, N’s were replaced by gaps. The PAGAN output was then analysed using “binary-table-by-edges-v3.1” (https://github.com/corydunnlab/hummingbird) (Dunn *et al.* 2019) to allow for a sum of substitutions at alignment positions encoded by human mtDNA. All fluctuating edges were extracted using “report-on-F-values-v1.1” (https://github.com/corydunnlab/hummingbird). Custom scripts were used to merge various data tables. Gap percentage for each position, with mammal and primate concatenates handled separately, was calculated using trimAl v1.4.rev22 (Capella-Gutiérrez *et al.* 2009). Shannon entropy was calculated using the ‘Shannon’ script (https://gist.github.com/jrjhealey/). Correspondence files for the human mtDNA (NC_012920.1) convention and alignment positions were generated using “extract-correspondence-for-merged-alignment-v.1.1.py”, followed by the construction of the look-up tables of amino acid direct substitution and presence using “direct-subst-lookup-table-proteins-v.1.1.py” and “AA-presence-lookup-table-v.1.1.py”, respectively (https://github.com/corydunnlab/Edge_mapping_conservation). All scripts were written using Python 2.7.

Chimera v1.13.1 was used for plotting structural data (Pettersen *et al.* 2004), and statistical testing and graph production were performed using Prism 8.3.0 (https://www.graphpad.com). Heat maps were generated using Matrix2png (Pavlidis and Noble 2003).

Database utilization

MITOMAP data used in this study (Lott *et al.* 2013) were downloaded on October 1, 2019. HelixMTdb data used in this study (Bolze *et al.* 2019) were downloaded on October 15, 2019.

Support vector machine development

Our SVM classifier (Cortes and Vapnik 1995) was developed using the R-language (https://www.R-project.org/) package e1071 (Meyer *et al.* 2014), whose implementation is based on libsvm (Chang and Lin 2011). A supervised binary classification model was built using positive and negative training sets. Each negative training set included alleles with the highest number of homoplasmic samples in HelixMTdb and also annotated as a polymorphism within MITOMAP. For proteins, the negative training sets consisted of 100 alleles, and for tRNAs, the negative training sets consisted of 50 alleles. The positive training sets for proteins consisted of alleles confirmed as pathogenic in MITOMAP. Development of a useful positive training set for tRNA pathogenicity required merging of the list of confirmed pathogenic tRNA alleles with a list of tRNA alleles reported (but not confirmed) to be pathogenic and manually selected based upon annotated symptoms or syndromes typically associated with mitochondrial disease. Analyzed positions for protein and tRNA included those gapped at 1.5% or less for mammalian or primate alignments. The best parameters for model fitting were found using the tune.svm wrapper in e1071. The best parameters for the radial basis kernel "rbfdot" were found by performing a $\log_2$ grid search of gamma and cost values during 5-fold cross-validation. Cross-validation error rate averaged from 10 independent repeats was used as the metric in choosing the best parameters. Subsequently, the full training set was used to train the SVM using these best parameters. Features were scaled internally in both training and when applying the model to predict the full data sets of mtDNA mutations. ROC curves were drawn with the mgraph function from rminer package (Cortez 2015), which uses ksvm from kernlab package (Karatzoglou *et al.* 2004). Feature importance was measured using the Importance function from rminer (Cortez and Embrechts 2013).

ACKNOWLEDGEMENTS

C.D.D. is funded by the Sigrid Jusélius Foundation and by an ERC Starter Grant (RevMito 637649). We thank Vivek Sharma and Gülayşe İnce Dunn for helpful comments on the manuscript.

FIGURE LEGENDS

Figure 1: A workflow for generation of SumSub(Int) values and substitution information from mtDNA sequence data. (**a**) Sequences for each macromolecule class (protein, rRNA, tRNA) are selected, aligned, and concatenated. (**b**) From these concatenates, phylogenetic trees are produced, followed by (**c**) ancestral prediction performed using alignments and phylogenetic trees as input. (**d**) Next, a binary matrix is generated, in which the presence or absence of mutations at each aligned position along each tree edge is tabulated. Conversion to a specific species convention is then achieved, if desired. (**e**) From this matrix, substitutions at internal tree edges, which are less subject to sequence errors, alignment errors, and unselected polymorphisms, can be summed. (**f**) Using the predicted internal node values, apparent direct substitution between ancestral and descendant characters are determined.

Figure 2: Most protein- and RNA-encoded positions in the human mitochondrial genome are under selective constraint. Concatenated sequences of each macromolecule class were analyzed at two different levels of taxonomy. Distributions of SumSub(Int) values are provided for all positions equal to or less than 1% gapped in their respective alignments: (**a**) mammalian proteins, (**b**) primate proteins, (**c**) mammalian rRNAs, (**d**) primate rRNAs, (**e**) mammalian tRNAs, (**f**) primate tRNAs.

Figure 3: SumSub(Int) values reflect conservation associated with important structural and catalytic residues. (**a**) Mitochondria-encoded subunits of the oxidative phosphorylation machinery often differ in the extent of positions under evolutionary constraint. Distributions of SumSub(Int) values for proteins analyzed using mammalian mtDNA sequence data were analyzed using a Kolmogorov-Smirnov test. Approximate P-value of < 0.0005, red; < 0.005, orange; < 0.05, yellow; >= 0.05, green. Kolmogorov-Smirnov D values are provided in (**b**). (**c**) Subunits of COX exhibit the most conservation, as reflected by SumSub(Int) values across positions. The same data examined above were analyzed by a box and whiskers plot, with whiskers indicating minimum and maximum values, box encompassing 25 through 75th percentile, and the line internal to the box representing the median. (**d**) Side chains of mtDNA-encoded COX subunits that are associated with high SumSub(Int) values face away from core, catalytic residues. Mitochondria-encoded subunits COI, COII, and COIII from one half of the dimeric crystal structure of *Bos taurus* cytochrome *c* oxidase (PDB: 1OCO) are represented in colored, ribbon format, with the remainder of the polypeptides depicted in grey, wire format. Bound heme groups are painted red and orange, and the conserved, catalytic E242 residue is colored green. SumSub(Int) values from analysis of mammalian phylogenetic data, without respect to gap percentage, are reflected as follows: 0 substitutions, black; 1-6 substitutions, dim grey; 7-12 substitutions, blue; 13-24 substitutions, cornflower blue; 25 substitutions and above, purple. In (**d**), the mitochondrial matrix lies at the bottom of the model. In (**e**), the viewer looks down toward the mitochondrial matrix.

Figure 4: SumSub(Int) values can report upon evolutionary divergence. (**a**) Degenerate positions characterized by the top 1% of SumSub(Int) values and gapped at 1% or less were extracted during

analysis of mammalian protein concatenates. The number of those degenerate sites mutated along each edge of the mammalian protein tree is plotted against the total number of substitutions along the tree edge. Superimposed upon these data is a restricted cubic spline curve. (**b**) Maximal SumSub(Int) values are greater for a tree encompassing mammals than for a tree comprised of primates. The top 2% of SumSub(Int) values at positions gapped 1% or less for each molecule class and for each clade are presented.

Figure 5: SumSub(Int) values provide unlimited dynamic range and can scale with level of taxonomy. SumSub(Int) values obtained from positional analysis of (**a**) protein, (**b**) rRNA, or (**c**) tRNA sequence concatenates and primate and mammal phylogenetic trees are compared to Shannon entropy values for the same alignment positions. All positions shown here are gapped at 1% or less.

Figure 6: Confirmed pathogenic mutations are associated with low SumSub(Int) values. The SumSub(Int) values for positions of alleles annotated in the MITOMAP database as either confirmed pathogenic or as polymorphisms are plotted for (**a**) protein positions analyzed using mammalian data, (**b**) protein positions analyzed using primate data, (**c**) tRNA positions analyzed using mammalian data, and (**d**) tRNA positions analyzed using primate data. All analyzed positions are gapped at equal to or less than 1%. ****, Kolmogorov-Smirnov test approximate P-value of <0.0001; ***, 0.0005. Dotted horizontal lines, median, with solid horizontal lines providing interquartile range.

Figure 7: Alleles currently annotated as polymorphic are deleterious. Alleles found within the MITOMAP database and encoding protein substitutions were classified based on whether an apparent direct substitution between the human reference character and the mutant character could be detected within a mammalian phylogenetic tree. All analyzed positions are gapped at equal to or less than 1%. Population prevalence for each allele is provided, with a trend toward low population frequency indirectly indicating pathogenicity. ****, Kolmogorov-Smirnov test approximate P-value of <0.0001.

Figure 8: A machine-learning approach predicts pathogenicity of mitochondrial polymorphisms in protein-coding genes using sequence data from mammals. (**a**) SVM-based prediction of pathogenicity for protein-coding alleles found in MITOMAP and HelixMTdb using features taking advantage of mammalian alignment positions of gap value 1.5% or less. Positive and negative training sets are superimposed on the curve of alleles and decision values. For clarity, the positive training set (**b**) and the negative training (**c**) set are also shown separately, demonstrating the lack of false positives within the predictive model. (**d**) Feature importance is calculated by calculation of the model's prediction error after shuffling its values. Error bars, mean with 95% confidence interval. (**e**) The receiver operating characteristic (ROC) curve demonstrates very high sensitivity is achieved with no loss in specificity for mammalian protein predictions. The true positive rate (sensitivity) of the predictive model is plotted against the false positive rate (1 - specificity) at various decision value thresholds. (**f**) Those MITOMAP alleles reported, but not confirmed to be associated with disease are plotted upon the prediction curve.

Figure 9: Protein-coding alleles predicted as pathogenic by a support vector machine differ in sample heteroplasmy and population frequency from those alleles predicted as non-pathogenic. (**a**) Protein-coding alleles predicted to be pathogenic are more commonly heteroplasmic than alleles predicted to be non-pathogenic. Alleles restricted to HelixMTdb were separated by SVM predictions

using mammal or primate sequences, and the fraction of population samples from each class for which the allele is always found to be heteroplasmic is shown. (**b**) The percentage of protein-coding alleles with detectable heteroplasmy represented within HelixMTdb is, following SVM analysis of mammalian data, higher for samples predicted to be pathogenic than for samples predicted to be non-pathogenic. Percent of samples heteroplasmic is plotted against alleles. (**c**) as in (**b**), but SVM learning was performed using primate data. (**d**) Pathogenicity predicted by an SVM approach is associated with a lower frequency of allele acquisition during clinical sampling. For the prediction sets examined above and generated using mammalian sequence data, the number of human samples harboring a given allele is plotted. (**e**) as in (**d**), but but SVM learning was performed using primate data. ****, Kolmogorov-Smirnov test approximate P-value of <0.0001.

SUPPLEMENTAL FIGURE LEGENDS

Figure S1: Selective constraint upon individual polypeptides as determined by analysis of mammalian protein alignments. Analysis of (**a**) Complex I, (**b**) Complex III, (**c**) Complex IV, and (**d**) Complex V proteins was performed as in Figure 2a.

Figure S2: Selective constraint upon individual polypeptides as determined by analysis of primate protein alignments. Analysis of (**a**) Complex I, (**b**) Complex III, (**c**) Complex IV, and (**d**) Complex V proteins was performed as in Figure 2b.

Figure S3: Selective constraint upon individual RNAs as determined by alignment analysis. (**a**) Analysis of mitochondria-encoded rRNAs performed using mammalian alignments as in Figure 2c. (**b**) Analysis of mitochondria-encoded rRNAs performed using primate alignments as in Figure 2d. (**c**) Analysis of tRNAs encoded upon the human mtDNA H-strand performed using mammalian alignments as in Figure 2e. (**d**) Analysis of tRNAs encoded upon the human mtDNA L-strand performed using mammalian alignments as in Figure 2e. (**e**) Analysis of tRNAs encoded upon the human mtDNA H-strand performed using primate alignments as in Figure 2f. (**f**) Analysis of tRNAs encoded upon the human mtDNA L-strand performed using mammalian alignments as in Figure 2f.

Figure S4: Comparison of total substitution number and summation of substitution at degenerate sites. Comparisons are carried out as in Figure 4b for (**a**) proteins analyzed using primate data, (**b**) rRNAs analyzed using mammal data, (**c**) rRNAs analyzed using primate data, (**d**) tRNAs analyzed using mammal data, and (**e**), tRNAs analyzed using primate data.

Figure S5: Apparent direct substitutions in Complex I proteins found along edges of a mammalian tree of mtDNA-encoded proteins. For each map, amino acid possibilities are provided on the X axis and amino acid positions within the human convention are listed along the Y-axis. White boxes include the human character and any amino acids reached by an apparent direct substitution along tree internal edges to or from the human character. Black boxes represent amino acids not found as acceptable direct substitutions from the human character within the same tree.

Figure S6: Apparent direct substitutions in Complex III, IV, and V proteins found along edges of a mammalian tree of mtDNA-encoded proteins. Calculations and panel preparation are performed as for Figure S5.

Figure S7: A machine-learning approach predicts pathogenicity of mitochondrial polymorphisms in protein-coding genes using sequence data from primates. As in Figure 8a - 8f, except primate protein data were utilized.

Figure S8: A machine-learning approach predicts pathogenicity of mitochondrial polymorphisms in tRNAs using sequence data from mammals. As in Figure 8a - 8e, except mammal tRNA data were utilized.

Figure S9: A machine-learning approach predicts pathogenicity of mitochondrial polymorphisms in tRNAs using sequence data from primates. As in Figure 8a - 8e, except primate tRNA data were utilized.

SUPPLEMENTAL TABLE LEGENDS

Table S1: Inferred direct amino acid substitutions at each human alignment position between the human reference amino acid and other amino acids are quantified within internal edges of a mammalian phylogenetic tree of mitochondria-encoded proteins. These changes were extracted from the appropriate PAGAN output using the script “direct-subst-lookup-table-proteins-v.1.1.py”.

Table S2: Inferred direct amino acid substitutions at each human alignment position between the human reference amino acid and other amino acids are quantified within internal edges of a primate phylogenetic tree of mitochondria-encoded proteins. Analysis was performed as in Table S1.

Table S3: Inferred direct amino acid substitutions at each human alignment position between the human reference amino acid and other amino acids are quantified within all edges of a mammalian phylogenetic tree of mitochondria-encoded proteins. Analysis was performed as in Table S1.

Table S4: Inferred direct amino acid substitutions at each human alignment position between the human reference amino acid and other amino acids are quantified within all edges of a primate phylogenetic tree of mitochondria-encoded proteins. Analysis was performed as in Table S1.

Table S5: Presence or absence of each amino acid at each human alignment position is inferred by study of internal edges of a mammalian phylogenetic tree of mitochondria-encoded proteins. These changes were extracted PAGAN output using the script “AA-presence-lookup-table-v.1.1.py”.

Table S6: Presence or absence of each amino acid at each human alignment position is inferred by study of internal edges of a primate phylogenetic tree of mitochondria-encoded proteins. Analysis was performed as in Table S5.

Table S7: Presence or absence of each amino acid at each human alignment position is inferred by study of all edges of a mammal phylogenetic tree of mitochondria-encoded proteins. Analysis was performed as in Table S5.

Table S8: Presence or absence of each amino acid at each human alignment position is inferred by study of all edges of a primate phylogenetic tree of mitochondria-encoded proteins. Analysis was performed as in Table S5.

Table S9: SVM output analyzing protein-encoding alleles found in the MITOMAP and HelixMTdb by use of mammalian sequence data. Details of SVM analyses and features are found in the methodology section.

Table S10: SVM output analyzing protein-encoding alleles found in the MITOMAP and HelixMTdb by use of primate sequence data. Details of SVM analyses and features are found in the methodology section.

Table S11: SVM output analyzing tRNA-encoding alleles found in the MITOMAP and HelixMTdb by use of mammalian sequence data. Details of SVM analyses and features are found in the methodology section.

Table S12: SVM output analyzing tRNA-encoding alleles found in the MITOMAP and HelixMTdb by use of primate sequence data. Details of SVM analyses and features are found in the methodology section.

REFERENCES

Alston C. L., M. C. Rocha, N. Z. Lax, D. M. Turnbull, and R. W. Taylor, 2017 The genetics and pathology of mitochondrial disease. J. Pathol. 241: 236–250.

Auer P. L., and G. Lettre, 2015 Rare variant association studies: considerations, challenges and opportunities. Genome Med. 7: 16.

Blell M., and M. A. Hunter, 2019 Direct-to-Consumer Genetic Testing's Red Herring: "Genetic Ancestry" and Personalized Medicine. Front. Med. 6: 48.

Bolze A., F. Mendez, S. White, F. Tanudjaja, M. Isaksson, *et al.*, 2019 Selective constraints and pathogenicity of mitochondrial DNA variants inferred from a novel database of 196,554 unrelated individuals. bioRxiv 1151.

Boulet L., G. Karpati, and E. A. Shoubridge, 1992 Distribution and threshold expression of the tRNA(Lys) mutation in skeletal muscle of patients with myoclonic epilepsy and ragged-red fibers (MERRF). Am. J. Hum. Genet. 51: 1187–1200.

Bratic A., and N.-G. Larsson, 2013 The role of mitochondria in aging. J. Clin. Invest. 123: 951–957.

Capella-Gutiérrez S., J. M. Silla-Martínez, and T. Gabaldón, 2009 trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972–1973.

Chamberlain S. A., and E. Szöcs, 2013 taxize: taxonomic search and retrieval in R. F1000Res. 2: 191.

Chang C.-C., and C.-J. Lin, 2011 LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST) 2: 27.

Chen L., P. Liu, T. C. Evans Jr, and L. M. Ettwiller, 2017 DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. Science 355: 752–756.

Chheda H., P. Palta, M. Pirinen, S. McCarthy, K. Walter, *et al.*, 2017 Whole-genome view of the consequences of a population bottleneck using 2926 genome sequences from Finland and United Kingdom. Eur. J. Hum. Genet. 25: 477–484.

Chinnery P. F., M. A. Johnson, T. M. Wardell, R. Singh-Kler, C. Hayes, *et al.*, 2000 The epidemiology of pathogenic mitochondrial DNA mutations. Ann. Neurol. 48: 188–193.

Chinnery P. F., and A. Gomez-Duran, 2018 Oldies but Goldies mtDNA Population Variants and Neurodegenerative Diseases. Front. Neurosci. 12: 883.

Cortes C., and V. Vapnik, 1995 Support-vector networks. Mach. Learn. 20: 273–297.

Cortez P., and M. J. Embrechts, 2013 Using sensitivity analysis and visualization techniques to open black box data mining models. Inf. Sci. 225: 1–17.

Cortez P., 2015 *A tutorial on using the rminer R package for data mining tasks*. Universidade do Minho. Escola de Engenharia (EEng).

Dayhoff M., R. Schwartz, and B. Orcutt, 1978 22 a model of evolutionary change in proteins, pp. 345–352 in *Atlas of protein sequence and structure*, National Biomedical Research Foundation Silver Spring MD.

Dunn C. D., B. A. Akpınar, and V. Sharma, 2019 An unusual amino acid substitution within hummingbird cytochrome c oxidase alters a key proton-conducting channel. bioRxiv 610915.

Elliott H. R., D. C. Samuels, J. A. Eden, C. L. Relton, and P. F. Chinnery, 2008 Pathogenic mitochondrial DNA mutations are common in the general population. Am. J. Hum. Genet. 83: 254–260.

Faith J. J., and D. D. Pollock, 2003 Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. Genetics 165: 735–745.

Fayet G., M. Jansson, D. Sternberg, A. R. Moslemi, P. Blondy, *et al.*, 2002 Ageing muscle: clonal expansions of mitochondrial DNA point mutations and deletions cause focal impairment of mitochondrial function. Neuromuscul. Disord. 12: 484–493.

Florentz C., B. Sohm, P. Tryoen-Tóth, J. Pütz, and M. Sissler, 2003 Human mitochondrial tRNAs in health and disease. Cell. Mol. Life Sci. 60: 1356–1375.

Fonseca R. R. da, W. E. Johnson, S. J. O'Brien, M. J. Ramos, and A. Antunes, 2008 The adaptive evolution of the mammalian mitochondrial genome. BMC Genomics 9: 119.

Goldman N., 1990 Maximum Likelihood Inference of Phylogenetic Trees, with Special Reference to a Poisson Process Model of DNA Substitution and to Parsimony Analyses. Syst. Biol. 39: 345–361.

Gorman G. S., A. M. Schaefer, Y. Ng, N. Gomez, E. L. Blakely, *et al.*, 2015 Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease. Ann. Neurol. 77: 753–759.

Greaves L. C., S. L. Preston, P. J. Tadrous, R. W. Taylor, M. J. Barron, *et al.*, 2006 Mitochondrial DNA mutations are established in human colonic stem cells, and mutated clones expand by crypt fission. Proc. Natl. Acad. Sci. U. S. A. 103: 714–719.

Hahn A., and S. Zuryn, 2019 The Cellular Mitochondrial Genome Landscape in Disease. Trends Cell Biol. 29: 227–240.

Hill G. E., J. C. Havird, D. B. Sloan, R. S. Burton, C. Greening, *et al.*, 2019 Assessing the fitness consequences of mitonuclear interactions in natural populations. Biol. Rev. Camb. Philos. Soc. 94: 1089–1104.

Iantorno S., K. Gori, N. Goldman, M. Gil, and C. Dessimoz, 2014 Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. Methods Mol. Biol. 1079: 59–73.

Jones D. T., W. R. Taylor, and J. M. Thornton, 1992 The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8: 275–282.

Karatzoglou A., A. Smola, K. Hornik, and A. Zeileis, 2004 kernlab-an S4 package for kernel methods in R. J. Stat. Softw. 11: 1–20.

Katoh K., and D. M. Standley, 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30: 772–780.

Kawrykow A., G. Roumanis, A. Kam, D. Kwak, C. Leung, *et al.*, 2012 Phylo: a citizen science approach for improving multiple sequence alignment. PLoS One 7: e31362.

Keinan A., and A. G. Clark, 2012 Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336: 740–743.

Kern A. D., and F. A. Kondrashov, 2004 Mechanisms and convergence of compensatory evolution in mammalian mitochondrial tRNAs. Nat. Genet. 36: 1207–1212.

Khan S., G. Ince-Dunn, A. Suomalainen, and L. L. Elo, 2020 Integrative omics approaches provide biological and clinical insights: examples from mitochondrial diseases. J. Clin. Invest. 130: 20–28.

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

Kumar S., 2005 Molecular clocks: four decades of evolution. Nat. Rev. Genet. 6: 654–662.

Landry L. G., N. Ali, D. R. Williams, H. L. Rehm, and V. L. Bonham, 2018 Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice. Health Aff. 37: 780–785.

Larsson A., 2014 AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30: 3276–3278.

Lott M. T., J. N. Leipzig, O. Derbeneva, H. M. Xie, D. Chalkia, *et al.*, 2013 mtDNA Variation and Analysis Using Mitomap and Mitomaster. Curr. Protoc. Bioinformatics 44: 1.23.1–26.

Löytynoja A., A. J. Vilella, and N. Goldman, 2012 Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. Bioinformatics 28: 1684–1691.

Luria S. E., and M. Delbrück, 1943 Mutations of Bacteria from Virus Sensitivity to Virus Resistance. Genetics 28: 491–511.

Maklakov A. A., L. Rowe, and U. Friberg, 2015 Why organisms age: Evolution of senescence under positive pleiotropy? Bioessays 37: 802–807.

Marom S., M. Friger, and D. Mishmar, 2017 MtDNA meta-analysis reveals both phenotype specificity and allele heterogeneity: a model for differential association. Sci. Rep. 7: 43449.

Matilainen O., P. M. Quirós, and J. Auwerx, 2017 Mitochondria and Epigenetics - Crosstalk in Homeostasis and Stress. Trends Cell Biol. 27: 453–463.

Medawar P. B., 1952 *An Unsolved Problem of Biology: An Inaugural Lecture Delivered at University College, London, 6 December, 1951*. H.K. Lewis and Company.

Meyer D., E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, *et al.*, 2014 e1071: Misc functions of the

Department of Statistics (e1071), TU Wien. R package version 1.

Nabholz B., H. Ellegren, and J. B. W. Wolf, 2013 High levels of gene expression explain the strong evolutionary constraint of mitochondrial protein-coding genes. Mol. Biol. Evol. 30: 272–284.

Nekhaeva E., N. D. Bodyak, Y. Kraytsberg, S. B. McGrath, N. J. Van Orsouw, *et al.*, 2002 Clonally expanded mtDNA point mutations are abundant in individual cells of human tissues. Proc. Natl. Acad. Sci. U. S. A. 99: 5521–5526.

Pavlidis P., and W. S. Noble, 2003 Matrix2png: a utility for visualizing matrix data. Bioinformatics 19: 295–296.

Pettersen E. F., T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, *et al.*, 2004 UCSF Chimera—a visualization system for exploratory research and analysis. J. Comput. Chem. 25: 1605–1612.

Phillips K. A., P. A. Deverka, G. W. Hooker, and M. P. Douglas, 2018 Genetic Test Availability And Spending: Where Are We Now? Where Are We Going? Health Aff. 37: 710–716.

Pozzi L., J. A. Hodgson, A. S. Burrell, K. N. Sterner, R. L. Raaum, *et al.*, 2014 Primate phylogenetic relationships and divergence dates inferred from complete mitochondrial genomes. Mol. Phylogenet. Evol. 75: 165–183.

Price M. N., P. S. Dehal, and A. P. Arkin, 2010 FastTree 2--approximately maximum-likelihood trees for large alignments, (A. F. Y. Poon, Ed.). PLoS One 5: e9490.

Raychaudhuri S., 2011 Mapping rare and common causal alleles for complex human diseases. Cell 147: 57–69.

Reis M. D., G. F. Gunnell, J. Barba-Montoya, A. Wilkins, Z. Yang, *et al.*, 2018 Using Phylogenomic Data to Explore the Effects of Relaxed Clocks and Calibration Strategies on Divergence Time Estimation: Primates as a Test Case. Syst. Biol. 67: 594–615.

Reyes A., C. Gissi, G. Pesole, and C. Saccone, 1998 Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol. Biol. Evol. 15: 957–966.

Shen W., S. Le, Y. Li, and F. Hu, 2016 SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS One 11: e0163962.

Smith D. R., 2019 Revisiting published genomes with fresh eyes and new data: Revising old sequencing data can yield unexpected insights and identify errors. EMBO Rep. 20: e49482.

Starr T. N., and J. W. Thornton, 2016 Epistasis in protein evolution. Protein Sci. 25: 1204–1218.

Stewart J. B., and P. F. Chinnery, 2015 The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. Nat. Rev. Genet. 16: 530–542.

Suomalainen A., and B. J. Battersby, 2018 Mitochondrial diseases: the contribution of organelle stress responses to pathology. Nat. Rev. Mol. Cell Biol. 19: 77–92.

Tanaka M., and T. Ozawa, 1994 Strand asymmetry in human mitochondrial dna mutations. Genomics 22: 327–335.

Tuppen H. A. L., E. L. Blakely, D. M. Turnbull, and R. W. Taylor, 2010 Mitochondrial DNA mutations and human disease. Biochim. Biophys. Acta 1797: 113–128.

Upham N. S., J. A. Esselstyn, and W. Jetz, 2019 Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. PLoS Biol. 17: e3000494.

Vento J. M., and B. Pappa, 2013 Genetic counseling in mitochondrial disease. Neurotherapeutics 10: 243–250.

Wei W., A. Gomez-Duran, G. Hudson, and P. F. Chinnery, 2017 Background sequence characteristics influence the occurrence and severity of disease-causing mtDNA mutations. PLoS Genet. 13: e1007126.

Whiffin N., E. Minikel, R. Walsh, A. H. O'Donnell-Luria, K. Karczewski, *et al.*, 2017 Using high-resolution variant frequencies to empower clinical genome interpretation. Genet. Med. 19: 1151–1158.

Wittenhagen L. M., and S. O. Kelley, 2003 Impact of disease-related mitochondrial mutations on tRNA structure and function. Trends Biochem. Sci. 28: 605–611.

Wolff J. N., E. D. Ladoukakis, J. A. Enríquez, and D. K. Dowling, 2014 Mitonuclear interactions: evolutionary consequences over multiple biological scales. Philos. Trans. R. Soc. Lond. B Biol. Sci. 369: 20130443.

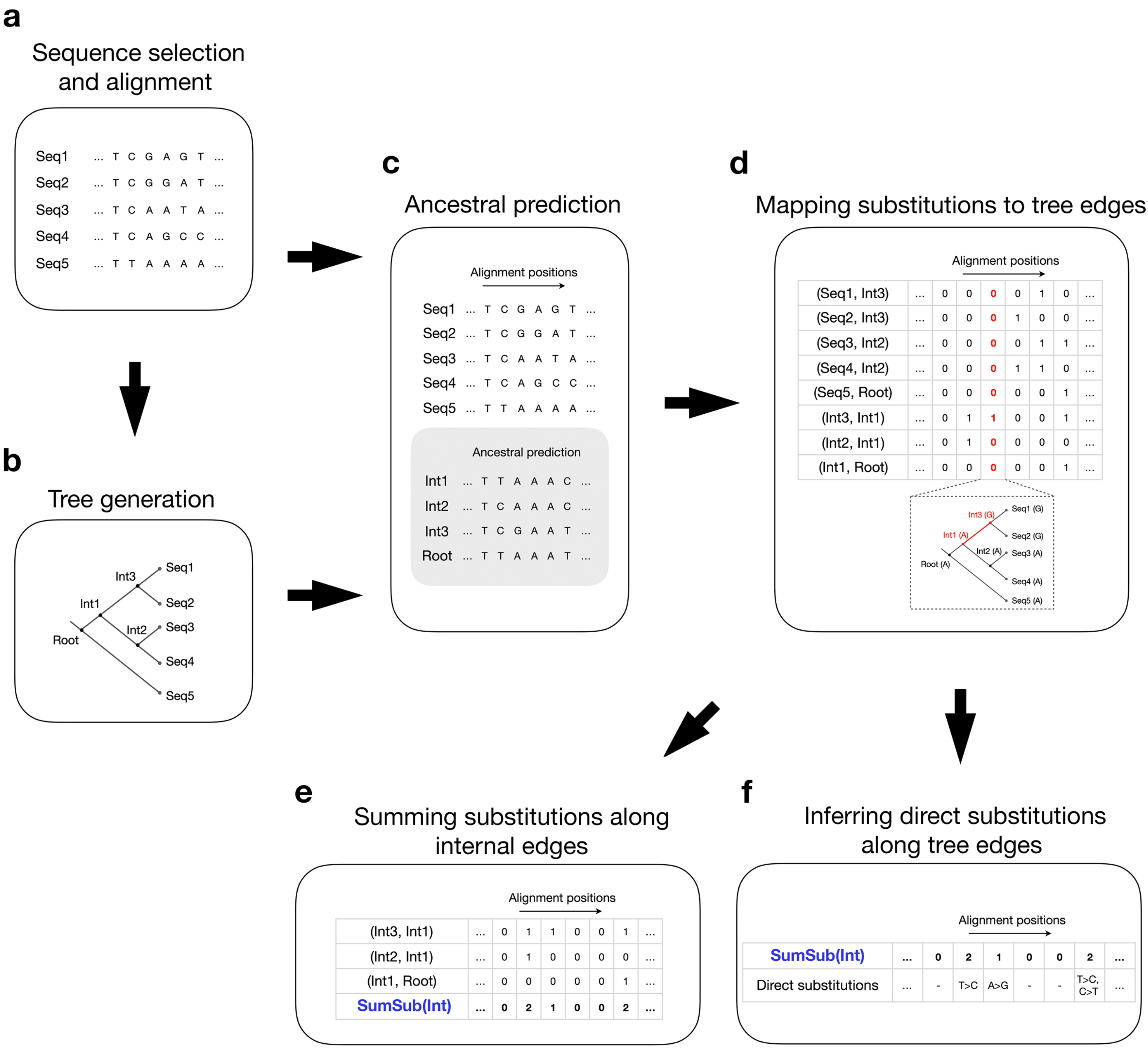
Yoshikawa S., K. Shinzawa-Itoh, R. Nakashima, R. Yaono, E. Yamashita, *et al.*, 1998 Redox-coupled crystal structural changes in bovine heart cytochrome c oxidase. Science 280: 1723–1729.

Zhang H., S. P. Burr, and P. F. Chinnery, 2018 The mitochondrial DNA genetic bottleneck: inheritance and beyond. Essays Biochem. 62: 225–234.

Zou Z., and J. Zhang, 2019 Amino acid exchangeabilities vary across the tree of life. Sci Adv 5: eaax3124.

Zuckerkandl E., and L. Pauling, 1965 Evolutionary Divergence and Convergence in Proteins. Evolving Genes and Proteins 97–166.

Zuk O., S. F. Schaffner, K. Samocha, R. Do, E. Hechter, *et al.*, 2014 Searching for missing heritability: designing rare variant association studies. Proc. Natl. Acad. Sci. U. S. A. 111: E455–64.



(Seq1, Int3)	...	0	0	0	0	1	0	...
(Seq2, Int3)	...	0	0	0	1	0	0	...
(Seq3, Int2)	...	0	0	0	0	1	1	...
(Seq4, Int2)	...	0	0	0	1	1	0	...
(Seq5, Root)	...	0	0	0	0	0	1	...
(Int3, Int1)	...	0	1	1	0	0	1	...
(Int2, Int1)	...	0	1	0	0	0	0	...
(Int1, Root)	...	0	0	0	0	0	1	...

(Int3, Int1)	...	0	1	1	0	0	1	...
(Int2, Int1)	...	0	1	0	0	0	0	...
(Int1, Root)	...	0	0	0	0	0	1	...
SumSub(Int)	**...**	**0**	**2**	**1**	**0**	**0**	**2**	**...**

SumSub(Int)	**...**	**0**	**2**	**1**	**0**	**0**	**2**	**...**
Direct substitutions	...	-	T>C	A>G	-	-	T>C, C>T	...

Figure 1

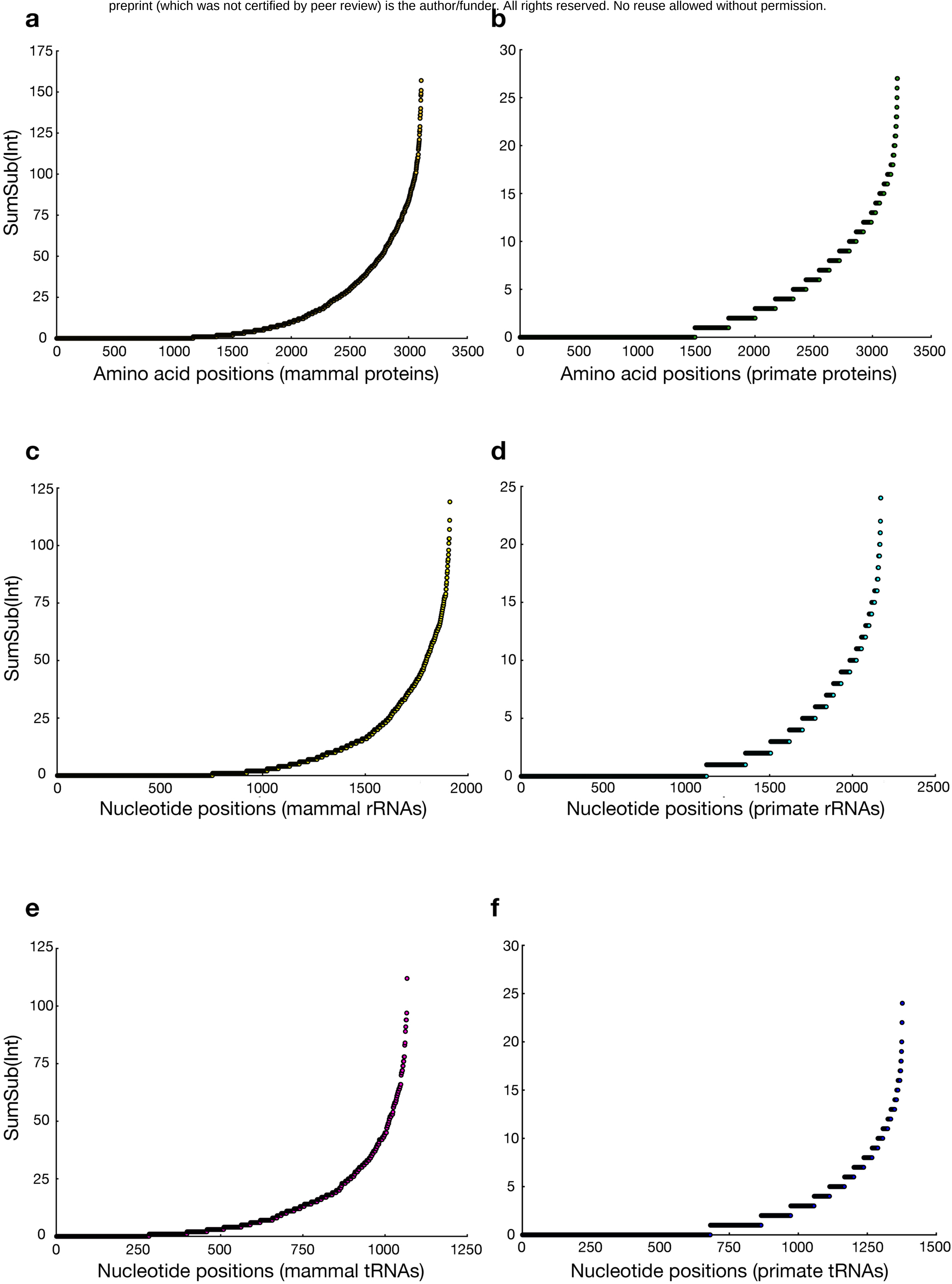

a
SumSub(Int)
0
25
50
75
100
125
150
175
0
500
1000
1500
2000
2500
3000
3500
Amino acid positions (mammal proteins)
b
0
5
10
15
20
25
30
0
500
1000
1500
2000
2500
3000
3500
Amino acid positions (primate proteins)
c
SumSub(Int)
0
25
50
75
100
125
0
500
1000
1500
2000
Nucleotide positions (mammal rRNAs)
d
0
5
10
15
20
25
0
500
1000
1500
2000
2500
Nucleotide positions (primate rRNAs)
e
SumSub(Int)
0
25
50
75
100
125
0
250
500
750
1000
1250
Nucleotide positions (mammal tRNAs)
f
0
5
10
15
20
25
30
0
250
500
750
1000
1250
1500
Nucleotide positions (primate tRNAs)

Figure 2

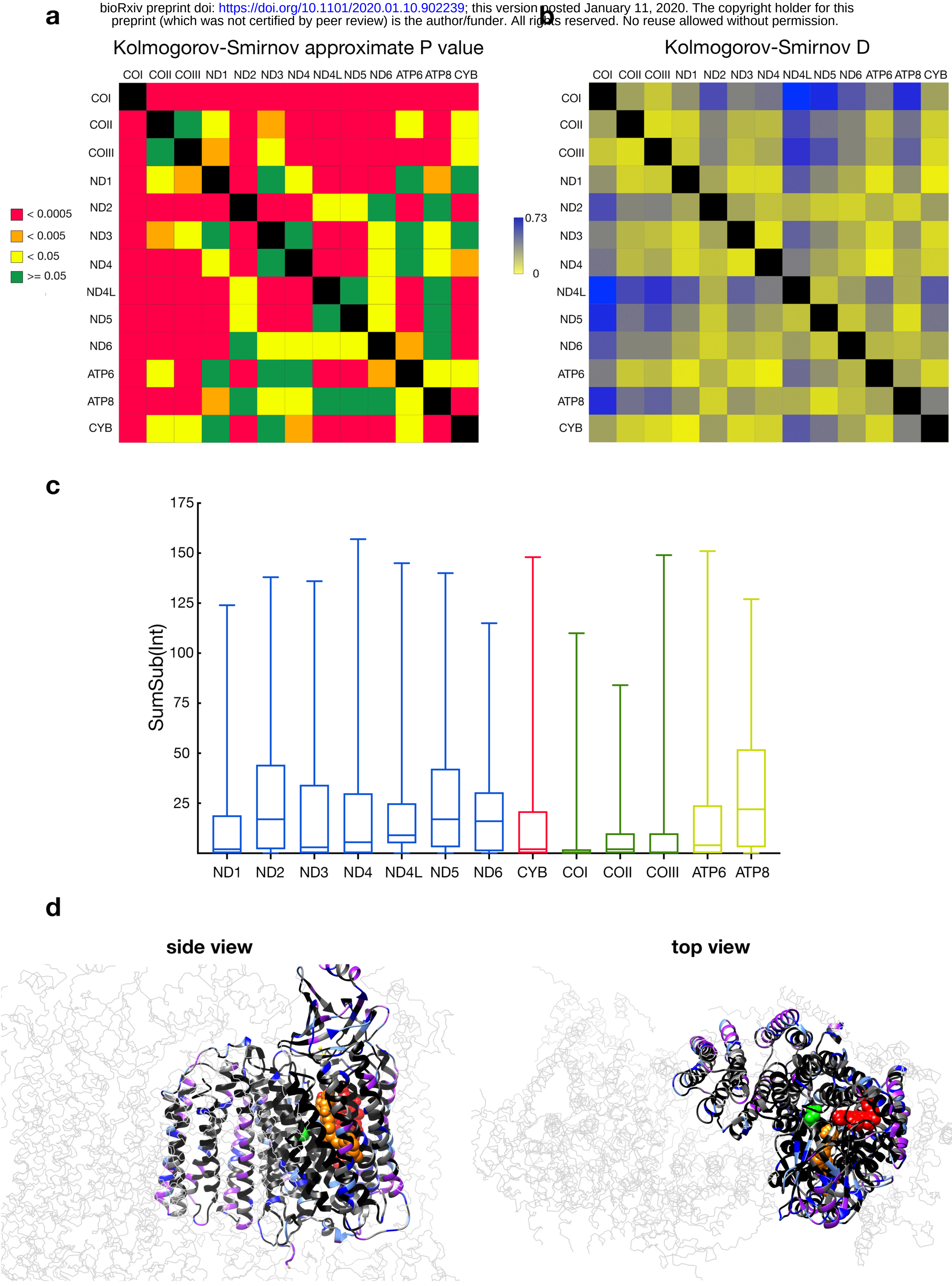

a
Kolmogorov-Smirnov approximate P value
COI COII COIII ND1 ND2 ND3 ND4 ND4L ND5 ND6 ATP6 ATP8 CYB
🟥 < 0.0005
🟧 < 0.005
🟨 < 0.05
🟩 >= 0.05
b
Kolmogorov-Smirnov D
COI COII COIII ND1 ND2 ND3 ND4 ND4L ND5 ND6 ATP6 ATP8 CYB
0.73
0
c
SumSub(Int)
175
150
125
100
75
50
25
ND1 ND2 ND3 ND4 ND4L ND5 ND6 CYB COI COII COIII ATP6 ATP8
d
side view
top view

Figure 3

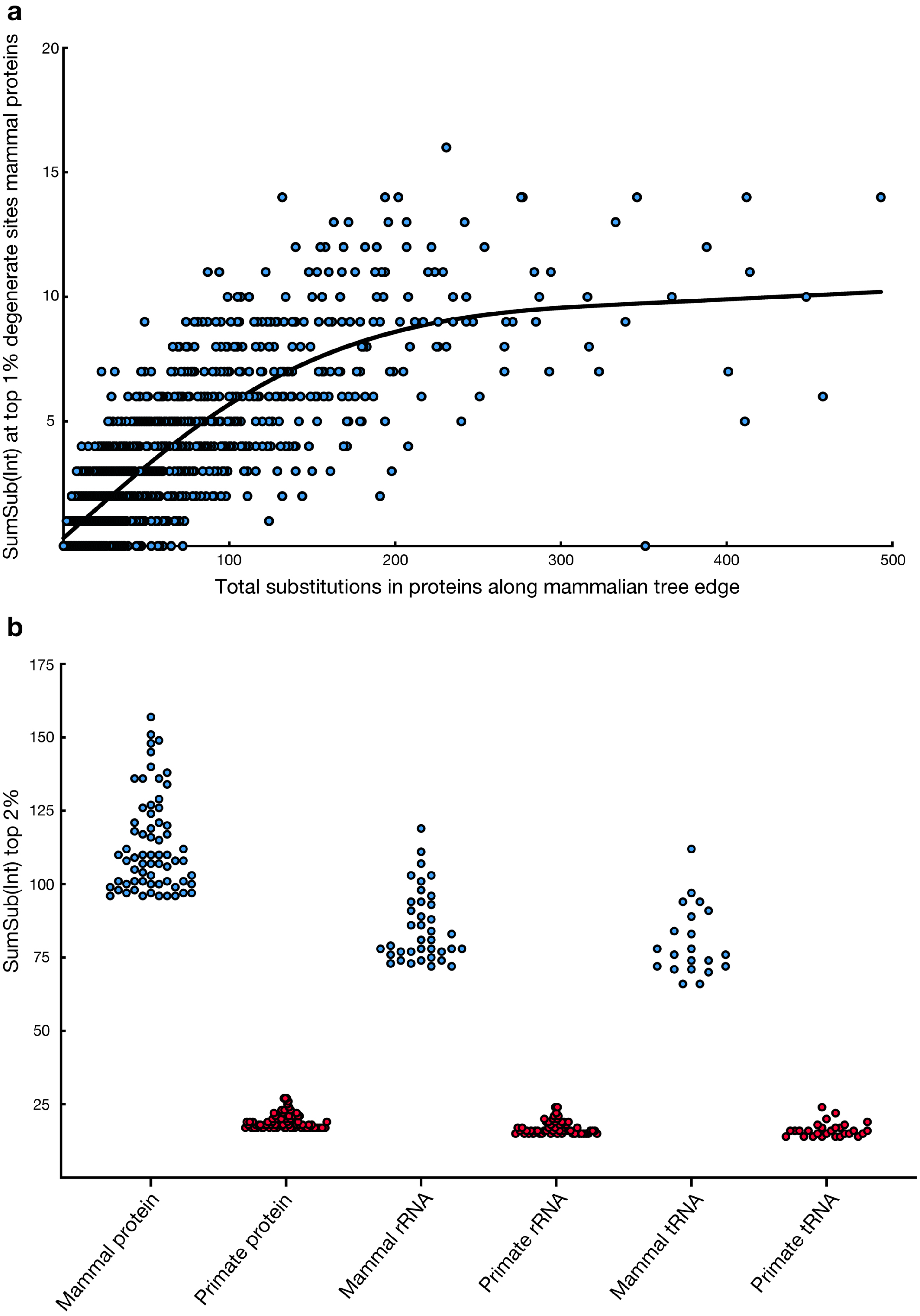

a
SumSub(Int) at top 1% degenerate sites mammal proteins
20
15
10
5
100
200
300
400
500
Total substitutions in proteins along mammalian tree edge
b
SumSub(Int) top 2%
175
150
125
100
75
50
25
Mammal protein
Primate protein
Mammal rRNA
Primate rRNA
Mammal tRNA
Primate tRNA

Figure 4

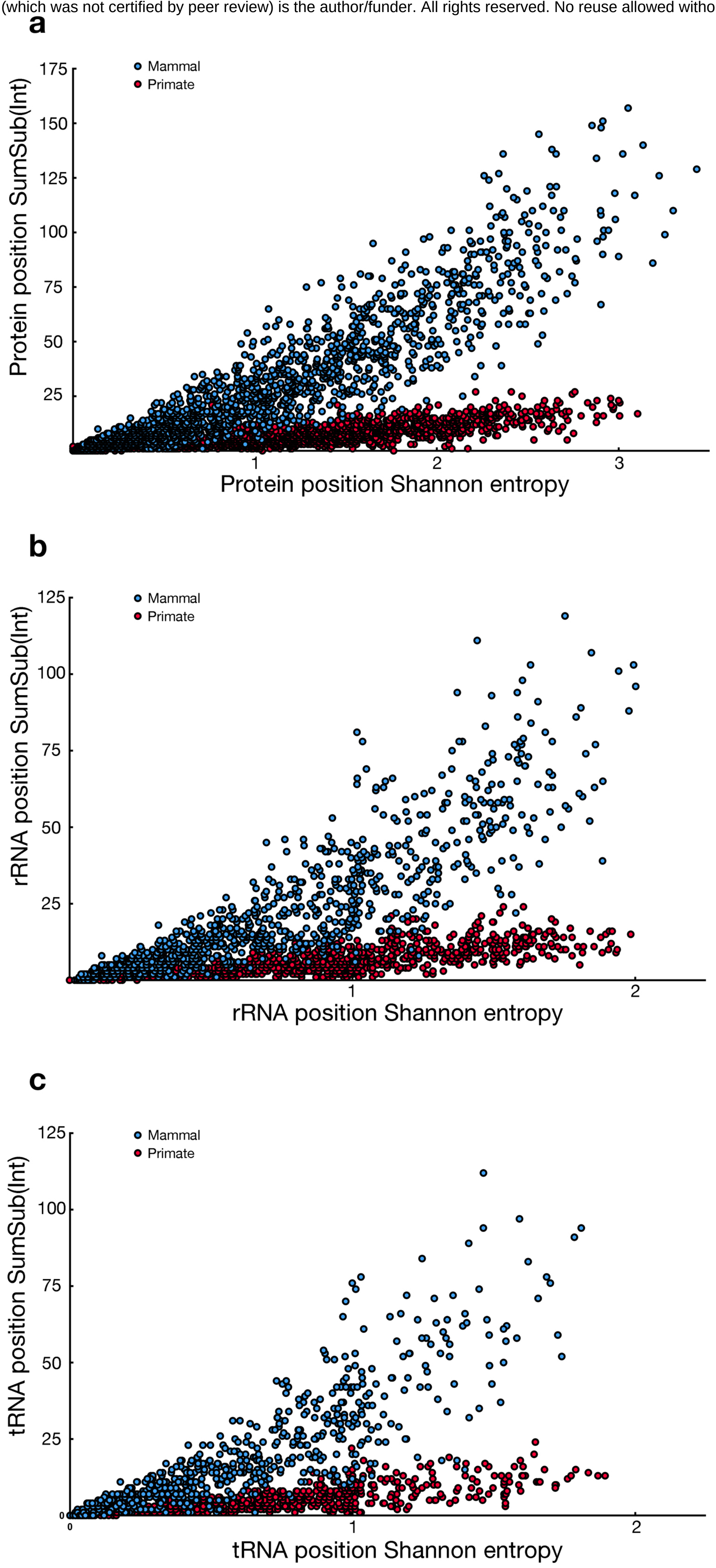

a
● Mammal
● Primate
Protein position SumSub(Int)
175
150
125
100
75
50
25
1
2
3
Protein position Shannon entropy
b
● Mammal
● Primate
rRNA position SumSub(Int)
125
100
75
50
25
1
2
rRNA position Shannon entropy
c
● Mammal
● Primate
tRNA position SumSub(Int)
125
100
75
50
25
0
0
1
2
tRNA position Shannon entropy

Figure 5

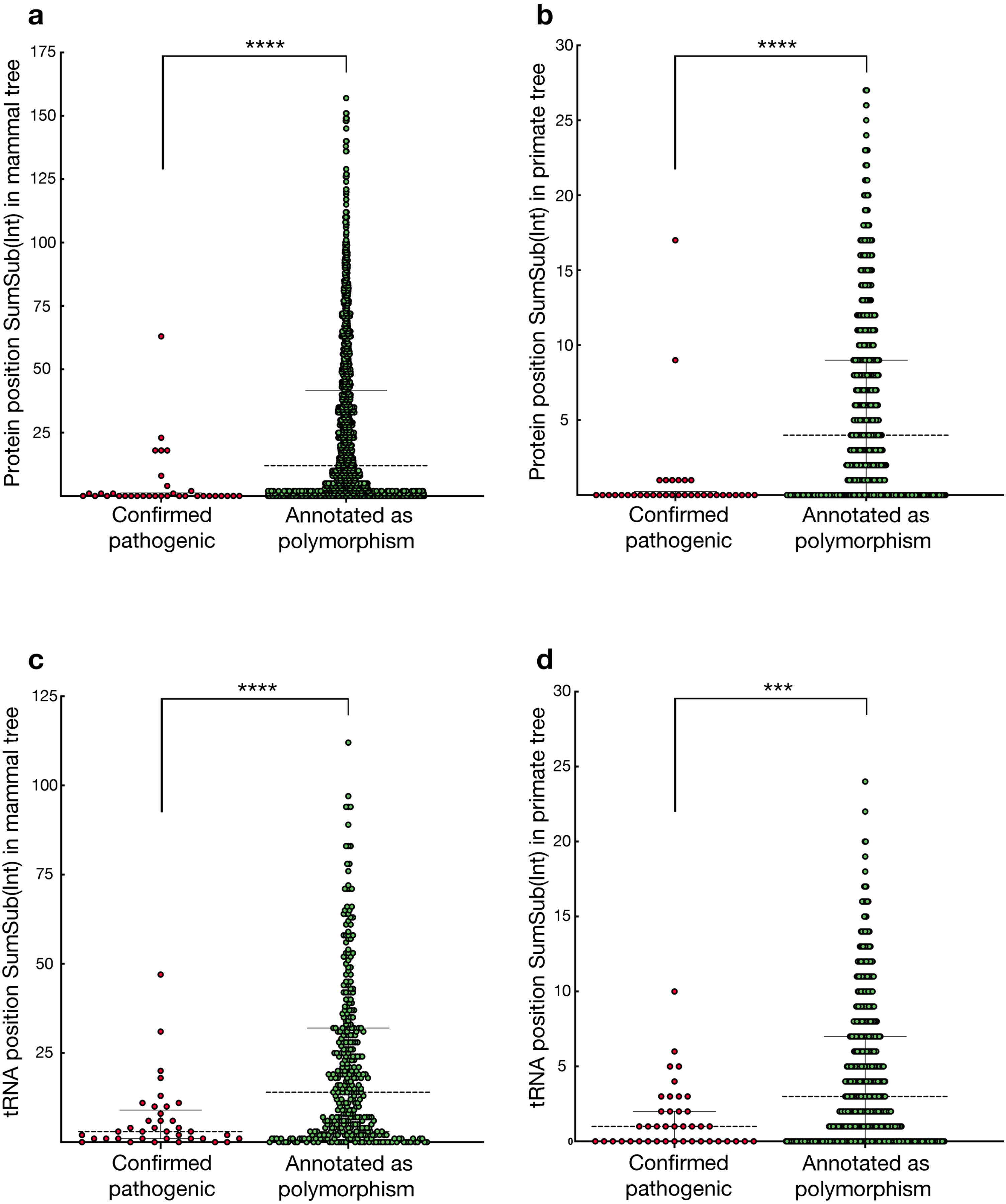

a
Protein position SumSub(Int) in mammal tree
175
150
125
100
75
50
25
****
Confirmed pathogenic
Annotated as polymorphism
b
Protein position SumSub(Int) in primate tree
30
25
20
15
10
5
****
Confirmed pathogenic
Annotated as polymorphism
c
tRNA position SumSub(Int) in mammal tree
125
100
75
50
25
****
Confirmed pathogenic
Annotated as polymorphism
d
tRNA position SumSub(Int) in primate tree
30
25
20
15
10
5
0
***
Confirmed pathogenic
Annotated as polymorphism

Figure 6

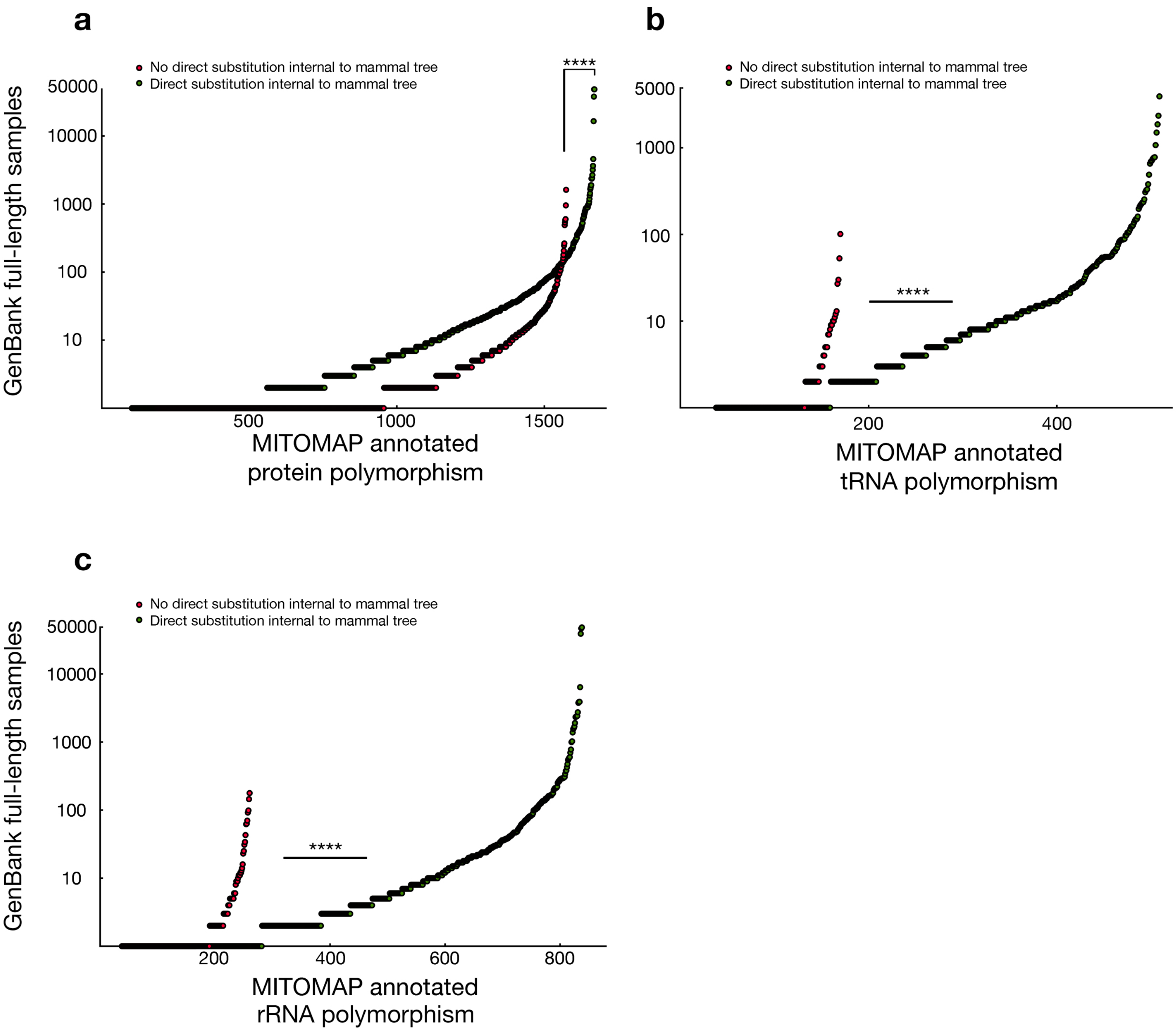

a
● No direct substitution internal to mammal tree
● Direct substitution internal to mammal tree
****
GenBank full-length samples
50000
10000
1000
100
10
500
1000
1500
MITOMAP annotated
protein polymorphism
b
● No direct substitution internal to mammal tree
● Direct substitution internal to mammal tree
5000
1000
100
10
****
200
400
MITOMAP annotated
tRNA polymorphism
c
● No direct substitution internal to mammal tree
● Direct substitution internal to mammal tree
GenBank full-length samples
50000
10000
1000
100
10
****
200
400
600
800
MITOMAP annotated
rRNA polymorphism

Figure 7

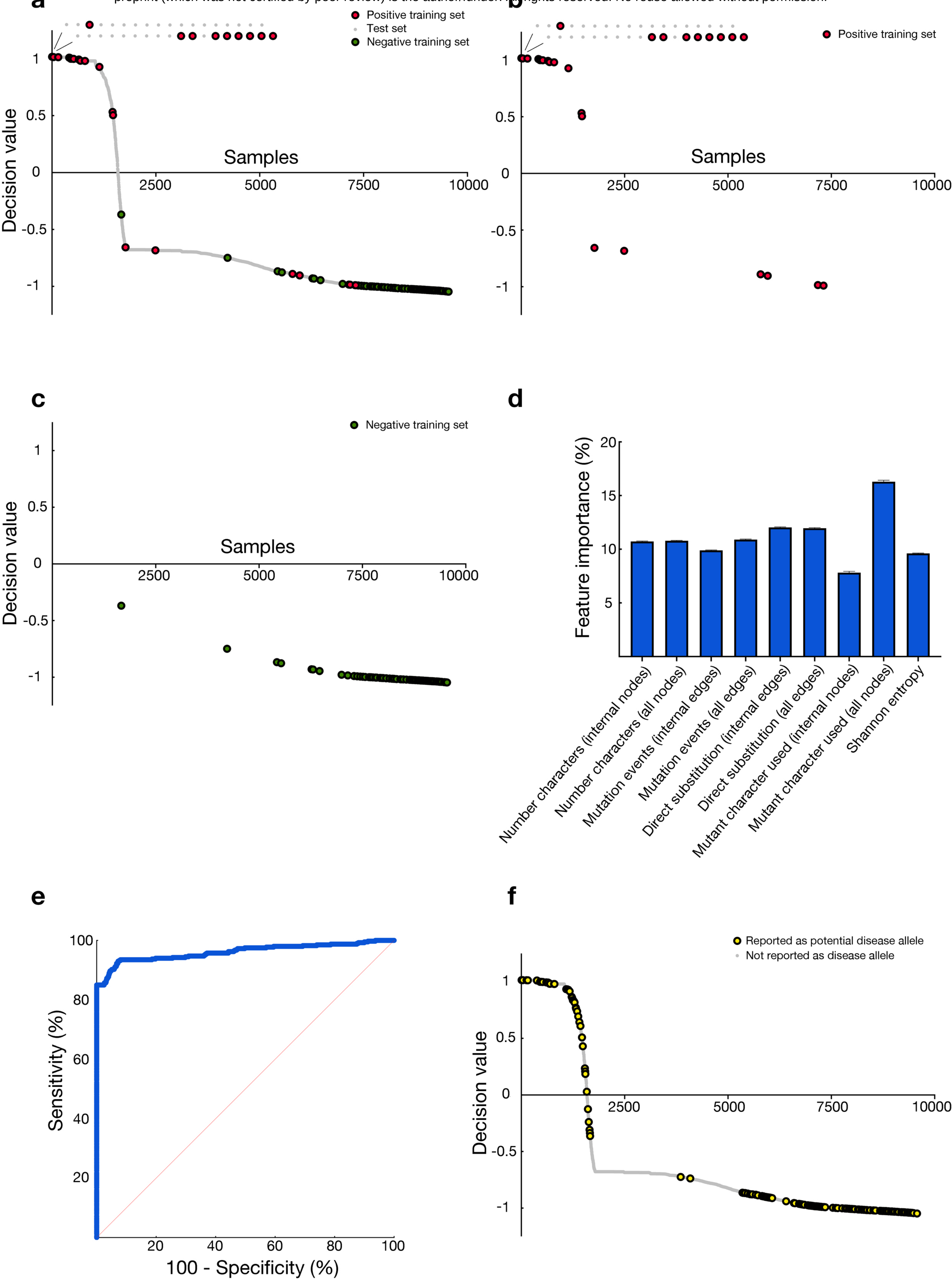

a
Positive training set
Test set
Negative training set
Decision value
1
0.5
0
-0.5
-1
Samples
2500
5000
7500
10000
b
Positive training set
1
0.5
0
-0.5
-1
Samples
2500
5000
7500
10000
c
Negative training set
Decision value
1
0.5
0
-0.5
-1
Samples
2500
5000
7500
10000
d
Feature importance (%)
20
15
10
5
Number characters (internal nodes)
Number characters (all nodes)
Mutation events (internal edges)
Mutation events (all edges)
Direct substitution (internal edges)
Direct substitution (all edges)
Mutant character used (internal nodes)
Mutant character used (all nodes)
Shannon entropy
e
Sensitivity (%)
100
80
60
40
20
100 - Specificity (%)
20
40
60
80
100
f
Reported as potential disease allele
Not reported as disease allele
Decision value
1
0.5
0
-0.5
-1
2500
5000
7500
10000

Figure 8

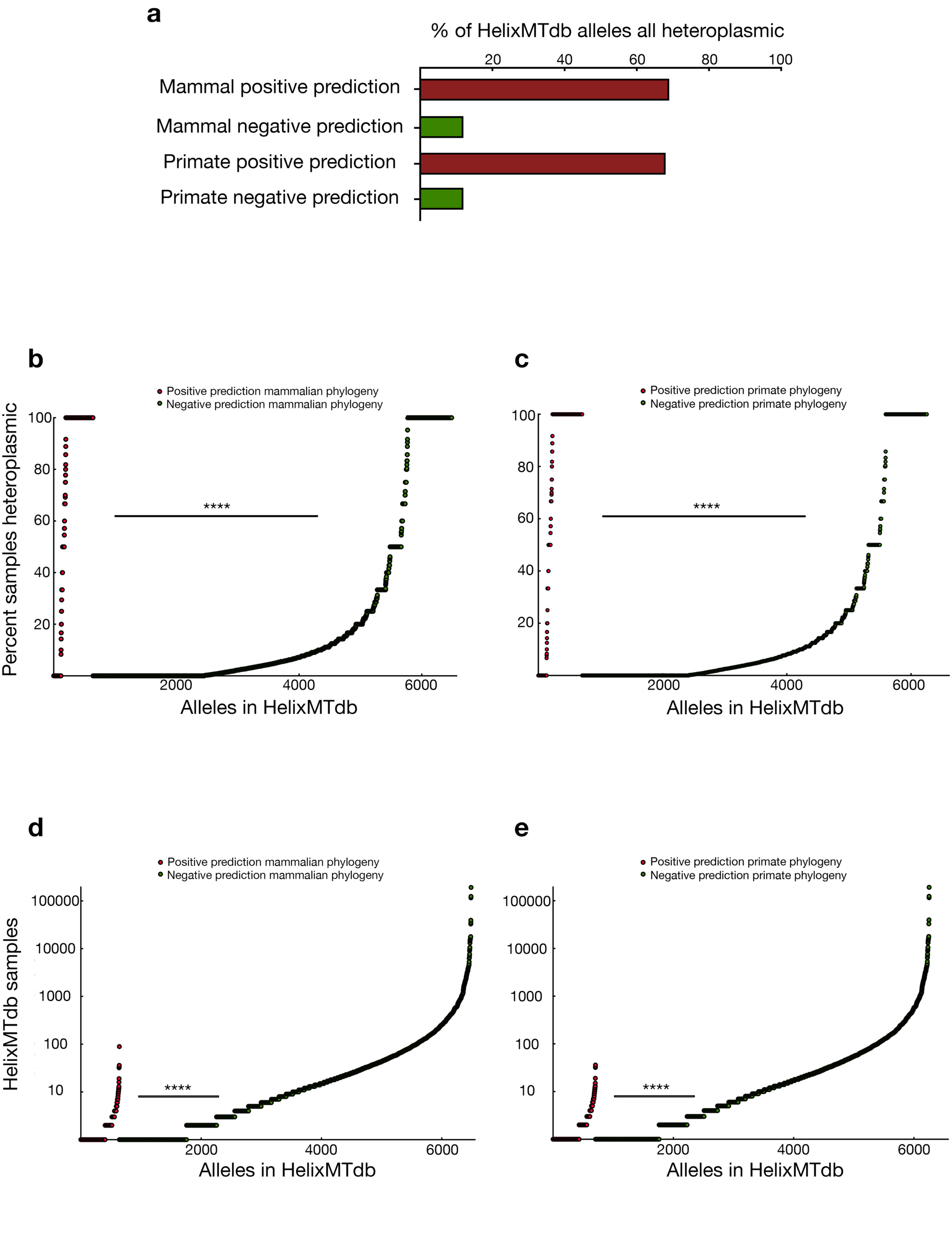

a
% of HelixMTdb alleles all heteroplasmic
20
40
60
80
100
Mammal positive prediction
Mammal negative prediction
Primate positive prediction
Primate negative prediction
b
● Positive prediction mammalian phylogeny
● Negative prediction mammalian phylogeny
Percent samples heteroplasmic
100
80
60
40
20
****
2000
4000
6000
Alleles in HelixMTdb
c
● Positive prediction primate phylogeny
● Negative prediction primate phylogeny
100
80
60
40
20
****
2000
4000
6000
Alleles in HelixMTdb
d
● Positive prediction mammalian phylogeny
● Negative prediction mammalian phylogeny
HelixMTdb samples
100000
10000
1000
100
10
****
2000
4000
6000
Alleles in HelixMTdb
e
● Positive prediction primate phylogeny
● Negative prediction primate phylogeny
100000
10000
1000
100
10
****
2000
4000
6000
Alleles in HelixMTdb

Figure 9