Pathogenicity of human mtDNA variants is revealed by combining a novel phylogenetic analysis with machine learning

Bala Anı Akpınar[1 †], Paul O. Carlson[1], Ville O. Paavilainen[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 remains a formidable 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 currently hamper clinical efforts to diagnose mitochondrial disease. Multiple sequence alignments are often deployed to estimate the potential significance of mitochondrial variants. However, factors including sample set bias, alignment errors, and sequencing errors currently limit the utility of multiple sequence alignments in pathogenicity prediction. Here, we describe an approach to assessment of site-specific conservation and variant acceptability that is reliant upon ancestral phylogenetic predictions and minimizes current alignment limitations. Using the output of our approach, we deploy machine learning in order to predict the pathogenicity of thousands of human mtDNA variants not yet linked to disease. Our work demonstrates that a substantial portion of mtDNA variants not yet characterized as harmful are, in fact, likely to be deleterious. Our findings will be of direct relevance to those at risk of mitochondria-associated illness, but the general applications of our methodology also extend beyond the context of mitochondrial disorders.

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 estimated to be overtly affected by mitochondrial disease. While a very limited number of mitochondrial DNA (mtDNA) lesions can be directly linked to human disease, the clinical outcome for many other mtDNA changes remains 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 different mtDNA alterations. 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 in the clinical setting, 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 our knowledge of mtDNA variants existing among humans continues to increase, comparison of which mtDNA variants are found among healthy or sick individuals (Whiffin *et al.* 2017) can 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 variant 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 variants that might act in a deleterious manner at the post-reproductive stage of life also makes likely the possibility that some mtDNA changes will contribute to age-related phenotypes while avoiding clear association with mitochondrial disease (Medawar 1952; Williams 1957; Maklakov *et al.* 2015; Cui *et al.* 2019).

Multiple sequence alignments offer important assistance when predicting a variant's potential pathogenicity (Raychaudhuri 2011). However, caveats are also associated with predicting mutation 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 lead to inaccurate clinical interpretations regarding the relative acceptability of specific variants (Keinan and Clark 2012; Zuk *et al.* 2014; Chheda *et al.* 2017; Landry *et al.* 2018). Third, interactions among alleles and environmental effects can

further complicate pathogenicity prediction (Kondrashov *et al.* 2002; Haut *et al.* 2004; Starr and Thornton 2016). Finally, alignment (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 substitution.

Here, we have deployed a novel methodology to calculate the relative conservation of human mtDNA-encoded positions with values of potentially unlimited dynamic range, and in a manner robust against sample bias, sequencing error, and alignment error. By subsequent application of machine learning, we demonstrate that a surprising number of uncharacterized mtDNA mutations carried by humans are likely to promote disease.

RESULTS

Mapping apparent substitutions to a phylogenetic tree allows calculation of relative positional conservation

We previously introduced an improved, empirical approach for detection and quantification of substitutions within evolving mtDNA-encoded macromolecules (Dunn *et al.* 2019). Here, we have built upon this approach and have applied our methodology toward prediction of human mitochondrial variant pathogenicity. Toward this goal, we first retrieved full mammalian mtDNA sequences from the National Center for Biotechnology Information Reference Sequence (NCBI RefSeq) database and extracted specific sequences of each RNA or protein-coding gene within the *Homo sapiens* reference mtDNA. Next, we aligned the resulting protein, tRNA, and rRNA sequences (Figure 1a), concatenated the sequences of each species based upon molecule class, and generated maximum likelihood trees (Figure 1b). Following tree generation, we performed ancestral prediction (Figure 1c) to reconstruct the character values of each position at every bifurcating node. Using the sequences of extant species and the predicted ancestral node values, we subsequently analysed each edge of the tree for the presence or absence of substitutions at each aligned human position (Figure 1d). Importantly, we focused 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 eventual misinterpretation 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. We subsequently sum the substitutions at internal positions to generate a new metric, the internal substitution score (ISS). The ISS should surpass metrics that consider positional character frequencies derived from multiple sequence alignments as a proxy of conservation, since character frequencies are highly sensitive to sample biases among extant sequences. Inferred ancestral and descendent characters at each edge were also extracted following generation of the substitution mapping matrix (Figure 1f), providing important information regarding which direct substitutions are possible within a lineage.

When used during analysis of macromolecules encoded by mammalian mtDNAs, it is clear that the ISS provides an excellent readout of relative conservation at, and consequent functional importance of, each alignment position. When plotting increasing ISSs across all amino acid positions with gap values of 1% or less, a characteristic curve was generated (Figure 2a). At the beginning of the curve, we encountered many positions that were nearly or completely invariant, implying critical roles for the folding or function of each polypeptide. Next, we observed a slope of increasing ISSs associated with amino acids of decreasing evolutionary constraint. The tail of each curve dramatically increases in slope, suggesting aligned, yet degenerate positions under minimal or no selection. Conservation curves are also obtained when studying individual mtDNA-encoded proteins within each oxidative phosphorylation complex (Figure 2b-e). Similar conservation curves are generated when examining combined (Figure 2f) and individual (Figure 2g) rRNA molecules or when considering combined (Figure 2h) and individual (Figure 2i-j) tRNA molecules that are encoded by the human mitochondrial genome.

When comparing ISS data from different mtDNA-encoded proteins, our methodology supports the idea 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 (da Fonseca *et al.* 2008; Nabholz *et al.* 2013) (Figure 3a). To further demonstrate that low ISSs correspond with conservation, we plotted mammalian ISS 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 ISSs extracted from the mammalian mtDNA tree, the most conserved positions were localized to the catalytic core of COX (Figure 3b-c and Movie S1). In contrast, more divergent amino acids clearly faced toward external regions of the complex, most prominent of which are the lipid-contacting regions of COX.

Internal substitution scores provide conservation measurements of unlimited dynamic range

Measures of conservation based upon character frequencies extracted from multiple sequence alignments, such as Shannon entropy, are inherently limited in dynamic range (Shannon 1948). Dynamic range limitations that can characterize frequency-based approaches to position analysis within alignments are especially prominent when considering nucleotides, where only four possibilities should occur at any given position. In contrast, summing positional substitutions across edges of a phylogenetic tree allows unlimited dynamic range of calculated conservation values. To demonstrate the increased dynamic range provided by our method, we compared Shannon entropy values and ISSs for well-aligned protein (Figure S1a), rRNA (Figure S1b), and tRNA (Figure S1c) positions encoded by the full set of mammalian mtDNAs and by a smaller set of mtDNAs limited to primates. Since Shannon entropy values and ISSs can both report upon conservation at aligned positions, these two variables were, as expected, correlated. However, while the range of Shannon entropy values cannot increase when moving from primate-level analysis to mammal-level analysis, the range of ISSs did increase,

suggesting that our approach to comparing selection at different sites benefits from the continuous accumulation of mtDNA sequence information.

We considered whether empirical measurements of substitution, rather than a model-based approach to conservation, would poorly reflect conservation if the evolutionary divergence between two nodes is 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. Next, we selected protein, rRNA, or tRNA sites that were under the most minimal selective constraint, as reflected by their rank in the top 1% of ISSs. We then compared the number of apparent changes summed across all positions (regardless of gap %) along an internal tree edge to the number of substitutions occurring only at these degenerate sites. The binary nature of our edge-focused substitution analysis, like other approaches investigating substitution, obscures the development of a Poisson distribution that would reflect repeated changes to degenerate sites (Zuckerkandl and Pauling 1965; Goldman 1990). Therefore, we expected that the total number of degenerate sites mutated along an edge would plateau beyond a given level of evolutionary divergence. 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 (Figure S2a). The relationship between rRNA (Figure S2b) and tRNA (Figure S2c) degenerate site substitutions and total edge substitutions is less clear-cut than than for proteins, potentially as a result of a lower number of degenerate positions available for analysis. However, a similar limitation on degenerate site substitution detection remains apparent beyond a given edge length. Since most positions encoded by mtDNA appear to be evolutionarily constrained and cannot reach the ISSs characterizing the most degenerate sites (Figure 2), and because most edges analyzed are shorter than the thresholds at which our plots become asymptotic, we expect that any potential under-counting of substitutions along longer tree edges will minimally affect our analysis of site-specific conservation.

Positional substitution frequency and inferred direct substitutions can be linked to human mtDNA variant pathogenicity

Since summation of detected substitutions across a phylogenetic tree provides a robust measure of relative conservation at different macromolecular positions, we were confident that a phylogenetic analysis that includes ISSs would be informative about the pathogenicity of human mtDNA variants. To test this possibility, we focused our attention upon substitutions annotated within the MITOMAP database of pathogenic mtDNA alterations (Lott *et al.* 2013). Indeed, we detected a clear relationship between confirmed pathogenicity and conservation at mtDNA-encoded amino acid positions, as reflected by positional ISSs (Figure 4a). Similarly, there was a strong link between tRNA mutation pathogenicity and ISSs obtained from mammalian tRNA data (Figure 4b). The paucity of confirmed pathogenic mitochondrial rRNA mutations in the MITOMAP database made comparisons using this class of molecules impractical, yet we expect future confirmation of additional pathogenic mutations in mitochondria-encoded rRNAs to permit future analyses. Together, findings obtained by phylogenetic analysis of mitochondria-encoded

proteins and tRNAs indicate that recovery of ISSs will be a valuable predictor of mtDNA mutations that lead to disease.

By ancestral prediction of nodes within our mammalian phylogenetic trees, we were able to infer ancestral and descendent characters along each tree edge. We reasoned that these data, while not totally uncoupled from positional ISSs, would provide additional information regarding what changes to mtDNA-encoded macromolecules may be harmful or harmless. Specifically, if direct substitution from the human reference character to the mutant character (or the inverse, assuming the time-reversibility of character substitutions) is predicted along the edge of a phylogenetic tree, then such a change at a given position might be expected to be less deleterious than an inferred direct substitution to or from the human character that was never encountered during evolution. In contrast, the simple presence or absence of a character at an alignment position taken out of context of its ancestor will often fail to reflect epistatic relationships that involve the position of interest (Kimura 1985; Kondrashov *et al.* 2002). We again focus upon internal tree edges, rather than those leading to extant sequences, so that we might avoid the confounding effects of sequencing and alignment errors.

To test the possibility that inferred internal direct substitutions (IDSs), like ISSs, may predict pathogenicity, we compared human variants confirmed to cause disease to those variants currently labelled as polymorphisms within the MITOMAP database. Pathogenic human amino acid changes from the reference sequence are rarely encountered among the internal edges of a mammalian phylogenetic tree (Figure 5a). In contrast, reportedly polymorphic alleles were much more likely to have the amino acid substitution identifiable within the mammalian phylogeny. For nucleotide substitutions in tRNAs, a similar analysis comparing confirmed pathogenic variants and variants annotated as polymorphisms detected no significant relationship between pathogenicity and IDSs. However, subsequent evidence indicated that IDSs in tRNAs may indeed help to differentiate harmful from harmless variants (see below).

Many mtDNA changes found in humans lead to uncharacterized pathogenicity

A subset of substitutions currently annotated as ‘polymorphic’ in the MITOMAP database were distinguished by very low ISS scores at the relevant alignment position, suggesting that mutations at those positions could, in fact, be deleterious. Moreover, when considering so-called polymorphisms, the absence of IDSs in mammals that matched human variations raised questions regarding whether additional pathogenicity may exist among human mtDNA substitutions currently considered polymorphic. In order to explore whether more pathogenic mutations might indeed be discovered among variants currently tabulated as ‘polymorphic,’ we tested whether protein polymorphisms tabulated in MITOMAP which cannot be directly linked to the human reference character in the mammalian phylogenetic tree would lead, through purifying selection, to a notable reduction in the number of full-length mtDNA samples found in GenBank. Measuring selection by assessment of population frequencies can be 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, the distribution of variant frequencies among full-length sequences in GenBank was strikingly different for those mutations for which an IDS could be identified in our mammalian trees of proteins (Figure 5b), tRNAs (Figure 5c), and rRNAs (Figure 5d) when compared to those for which an IDS could not be identified. These results indicate that many variants currently considered to be polymorphic are, in fact, harmful.

A support vector machine can predict deleterious mtDNA polymorphisms

Given the clear presence of deleterious substitutions among uncharacterized variants, we sought a high-throughput method that would reveal these potentially pathogenic changes. We turned to machine learning, and specifically a support vector machine (SVM) approach (Cortes and Vapnik 1995), to predict the risk of mtDNA variants in a manner dependent upon our phylogenetic analyses. Along with the mutations tabulated in MITOMAP, we also included mtDNA changes found in the recently established HelixMT database (HelixMTdb). This database was generated during a genetic survey, performed agnostic to phenotype, of nearly 200,000 independent human samples (Bolze *et al.* 2019). Positive training sets for our SVM consisted of protein or tRNA mutations listed as confirmed pathogenic alterations within the MITOMAP database. Negative training sets included those variants both annotated as polymorphic in MITOMAP and characterized by the highest counts of homoplasmic alleles within 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 ISS scores, but expanded our features to include the Total Substitution Score (TSS), 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 ever found in internal nodes or all nodes; and positional Shannon entropy.

For variants found within protein-coding genes, our SVM approach clearly permitted a robust and automated prediction of which polymorphisms not yet classified as pathogenic may cause or contribute to disease (Figure 6a and Table S1). Importantly, when examining the behavior of alleles within our negative training sets, no false-positives were predicted. Few confirmed pathogenic mutations were found among the predicted negative set. No specific feature, all of which are related to site conservation in mammals, stood out as dominant in providing predictive value (Figure 6b). 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 associated Receiver Operating Characteristic (ROC) curve (Figure 6c).

For tRNAs, a SVM-based approach again provided notable separation between positive and negative training sets (Figure 6d and Table S2). During tRNA analysis, the TSS and the number of characters present at internal nodes appeared to play the strongest role in prediction (Figure 6e). The ROC curve for our tRNA analysis suggests noteworthy predictive proficiency of the SVM approach for tRNAs, yet predictive power was more modest than that obtained during analysis of coding variants (Figure 6f) and the very strict delineation between positive and negative predictions that characterized our analysis of protein-coding positions was not

apparent. Due to the lack of confirmed pathogenic mutations in rRNA, we have not yet applied our predictive approach to this class of mtDNA-encoded macromolecules.

Beyond variants of confirmed pathogenicity used to train our SVM models, there are other changes provided in the MITOMAP database that are 'reported', yet not confirmed, to be pathogenic. We asked whether our SVM predictions might provide further evidence that some of these variants reported to be mitochondrial disease-associated are truly pathogenic. Toward this goal, we removed substitutions potentially related to carcinogenesis, then assigned SVM prediction values to the remaining variants. We expected that pathogenic alterations are more likely to be detected as heteroplasmic within human samples, since overt mitochondrial disease can require a deleterious variant to rise to a high proportion of the mtDNA molecules maintained by the cell (Stewart and Chinnery 2015; Hahn and Zuryn 2019). Therefore, we asked whether variants in our two predicted classes could be differentiated by the percentage of samples reported as heteroplasmic within HelixMTdb. Indeed, there was a clear difference in propensity to heteroplasmy between the two classes when considering mutations affecting protein (Figure 7a) and tRNA (Figure 7b). With a continued focus on MITOMAP variants reported, but not confirmed, to be disease-associated, we next assessed purifying selection by assessing the number of total samples in HelixMTdb for each variant within our prediction classes. Indeed, there was a decrease in population prevalence among protein-coding (Figure 7c) and tRNA (Figure 7d) variants that we had predicted as pathogenic. These findings demonstrate the general applicability of our pathogenicity predictions, but also provide additional evidence supporting the pathogenicity of specific mtDNA changes not yet fully confirmed as disease-causing within MITOMAP.

Following validation of our SVM prediction approach using a subset of variants enriched for pathogenicity, we turned to the thousands of mtDNA changes tabulated within MITOMAP or HelixMTdb that are not currently characterized as harmful. Coupling our evolutionary analysis of mtDNA-encoded macromolecules to machine learning again revealed the potential pathogenicity of hundreds of variants found within the human population. The heteroplasmy profile (Figure 7e) and population distribution (Figure 7f) were again clearly distinct when comparing those variants that we predict to be harmful and those variants that we predict to be benign. Similarly, although less prominently, tRNAs predicted to be harmful or harmless were also distinguished by heteroplasmy distribution (Figure 7g) and population frequency distribution (Figure 7h).

DISCUSSION

We have described here a methodology that allows improved quantification of the relative conservation of sites within and between genes, RNAs, and proteins. 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, will 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.

Beyond the development of a novel quantitative measure of conservation, we also use ancestral predictions to generate site-specific substitution matrices. 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 when moving across clades (Zou and Zhang 2019), and successful substitution of any given character 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.

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 mitochondrial disease. We note that our approach is not limited to the study of mitochondria-encoded components, but can be applied to any macromolecule for which sufficient sequence data exists toward generation of a high quality alignment, a well-supported phylogenetic tree, and reliable ancestral predictions.

Protein and tRNA variants of confirmed pathogenicity were clearly characterized by low ISSs. Furthermore, a lack of IDSs mimicking protein-coding mutations from the human reference sequence could also be linked to pathogenicity. But how should one weigh these and other available factors when classifying, in a high-throughput manner, the many variants already encountered or waiting to be discovered within human mtDNAs? To address this question, we deployed a machine learning approach, trained on existing data present in the MITOMAP database, toward determination of mtDNA mutation pathogenicity. Our methodology appeared to be very effective. Even though population frequency is subject to bias (see below), variants predicted to be pathogenic were characterized by low prominence in the population, suggesting purifying selection. Moreover, we detected a strong bias against homoplasmy when considering polypeptide mutations that we predicted to be harmful, a common indicator of pathogenicity.

mtDNA variants predicted to be pathogenic 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 substitutions remain to be classified as pathogenic. So why have these potentially harmful changes not yet been classified as harmful within the clinic?

First, if a mutation is too common, clinicians may quickly consider the mtDNA variant to be unlinked to patient symptoms (Whiffin *et al.* 2017). Our findings do, in fact, support the idea that population frequency of an alteration can correspond with pathogenicity. However, population frequency cannot be a dependable predictor of clinical outcome, since population

counts require substantial correction 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 variants 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 substitutions that we predict to be pathogenic, but are not yet clinically annotated as disease-associated. Differential distribution of mtDNA variants within a potential patient, either established during development or resulting from bottleneck effects in renewable and non-mitotic 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 variants 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 was then analysed using “binary-table-by-edges-v2.2” and "add-convention-to-binarytable-v1.1.py" (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). 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). For IDS determination, only edges with standard characters at both nodes were considered. For ISS calculation, 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. For positions that contained only N’s and gaps, N’s were replaced by gaps. Replacement of ambiguous characters with standard characters was carried out using using script “replace-nonstandard.py” (https://github.com/corydunnlab/Edge_mapping_conservation), and 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. 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/). Custom scripts were used to merge various data tables. Software generated within the context of this study was 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). When plotting data on a logarithmic scale, variant population counts of zero were converted to one.

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 variants, and for tRNAs, the negative training sets consisted of 50 variants. The positive training sets for proteins consisted of variants confirmed as pathogenic in MITOMAP. 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 during training and when applying the model in order 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.

DISCLOSURES

C.D.D., B.A.A., and P.O.C. are members of Primal Predictions LLC, a firm developing novel approaches to variant pathogenicity prediction.

FIGURE LEGENDS

Figure 1: A workflow for generation of internal substitution scores and inferred internal direct substitutions. (**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 with reference to a specific species convention. (**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, inferred direct substitutions between ancestral and descendant characters are determined.

Figure 2: Most well-aligned positions encoded by the human mitochondrial genome are under selective constraint. Sequences of each macromolecule class were analyzed as described in Figure 1. Distributions of ISSs are provided for all alignment positions equal to or less than 1% gapped in their respective alignments of mammalian sequences: (**a**) all mtDNA-encoded proteins considered together as a concatenate, (**b**) individual Complex I proteins, (**c**) individual Complex III proteins, (**d**) individual Complex IV proteins, (**e**) individual Complex V proteins, (**f**) concatenated, mitochondria-encoded rRNAs, (**g**) individual rRNAs, (**h**) concatenated, mtDNA-encoded tRNAs, (**i**) individual tRNAs encoded on the human H-strand, (**j**) individual tRNAs encoded on the human L-strand.

Figure 3: Internal substitution scores reflect conservation associated with important structural and catalytic residues. (**a**) Subunits of COX exhibit the most conservation, as reflected by ISSs 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. (**b-c**) Side chains of mtDNA-encoded COX subunits that are associated with high ISSs 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. ISSs 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 (**b**), the mitochondrial matrix lies at the bottom of the model. In (**c**), the viewer looks down toward the mitochondrial matrix.

Figure 4: Confirmed pathogenic mutations are associated with low internal substitution scores. ISSs for positions of variants annotated in the MITOMAP database as either confirmed pathogenic or as polymorphisms are plotted for (**a**) protein positions analyzed using mammalian data and (**b**) tRNA positions analyzed using mammalian data. All analyzed positions are gapped

at equal to or less than 1%. ****, Kolmogorov-Smirnov test approximate P-value of <0.0001. Dotted horizontal lines, median, with solid horizontal lines providing interquartile range.

Figure 5: Analysis of inferred internal direct substitutions indicates that variants currently annotated as polymorphic may be deleterious. (**a**) Inferred internal direct substitutions within the mammalian phylogenetic tree correspond to a lack of confirmed pathogenicity for mitochondrial protein coding variants. Confirmed pathogenic protein substitutions and substitutions currently annotated as polymorphisms, both retrieved from MITOMAP, were tested for the presence of an IDS. ****, Fisher's two-sided exact test P-value of <0.0001. All analyzed positions are gapped at equal to or less than 1%. (**b-d**) Population frequency distributions differ when MITOMAP-annotated substitutions from the human reference character in protein (**b**), tRNAs (**c**), or rRNAs (**d**) are classified based upon the presence or absence of an IDS. Variants found within the MITOMAP database and encoding protein substitutions were classified based on whether an IDS between the human reference character and the mutant (**b**) protein, (**c**) tRNA, or (**d**) rRNA character could be detected within a mammalian phylogenetic tree. All analyzed positions are gapped at equal to or less than 1%. ****, Kolmogorov-Smirnov test approximate P-value of <0.0001.

Figure 6: A machine-learning approach predicts pathogenicity of mtDNA variants currently annotated as polymorphisms. (**a**) SVM-based prediction of pathogenicity for protein-coding variants 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 variants and decision values. (**b**) Feature importance is calculated from the increase of the model's prediction error after shuffling the values for each feature. 10 test iterations are shown. (**c**) 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. (**d**) As for (**a**), but for tRNAs. (**e**) As for (**b**), but for tRNAs. (**f**) As for (**c**), but for tRNAs.

Figure 7: SVM-based pathogenicity prediction rooted in evolutionary analyses is supported by separable distributions of heteroplasmy profiles and allelic population frequencies. (**a**) For mitochondria-encoded protein substitutions reported to be disease-associated within the MITOMAP database, but not confirmed as pathogenic, the percent of HelixMTdb samples for a given variant that were heteroplasmic is plotted according to pathogenicity prediction by SVM. (**b**) As for (**a**), except tRNA variants are plotted. (**c**) For protein substitutions reported to be disease-associated within the MITOMAP database, but not confirmed as pathogenic, the total number of HelixMTdb samples for a given variant is plotted according to pathogenicity prediction by SVM. (**d**) As for (**c**), except tRNA substitutions are plotted. Many mitochondrial substitutions currently annotated as polymorphic are, in fact, pathogenic. (**e**) For protein substitutions in MITOMAP and HelixMTdb not associated with disease, the percent of heteroplasmic HelixMTdb samples for a given variant is plotted according to pathogenicity prediction. (**f**) For protein substitutions in MITOMAP and HelixMTdb not associated with disease, the number of samples with the given variant is displayed

according to pathogenicity prediction. (**g**) As in (**e**), except that tRNA changes are analyzed. (**h**) As in (**f**), except that tRNA substitutions are considered. ****, Kolmogorov-Smirnov test approximate P-value of <0.0001.

SUPPLEMENTAL FIGURE LEGENDS

Figure S1: Internal substitution scores provide unlimited dynamic range and can scale with level of taxonomy. ISSs 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 S2: Most edges from trees of mammalian mtDNA-encoded macromolecules are not saturated for substitutions. Degenerate positions characterized by the top 1% of ISSs and gapped at 1% or less were extracted during analysis of mammalian macromolecule 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. Analysis was performed for (a) proteins (b) rRNAs and (c) tRNAs.

SUPPLEMENTAL TABLE LEGENDS

Table S1: SVM output analyzing protein-encoding variants 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 S2: SVM output analyzing tRNA-encoding variants found in the MITOMAP and HelixMTdb by use of mammalian 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).

Cui R., T. Medeiros, D. Willemsen, L. N. M. Iasi, G. E. Collier, *et al.*, 2019 Relaxed Selection

Limits Lifespan by Increasing Mutation Load. *Cell* 178: 385–399.e20.

Dayhoff M., R. Schwartz, and B. Orcutt, 1978 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.

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.

Haut S., T. B. de Villemeur, M. Brivet, A. Guiochon-Mantel, A. Boutron, *et al.*, 2004 The deleterious G15498A mutation in mitochondrial DNA-encoded cytochrome b may remain clinically silent in homoplasmic carriers. *Eur. J. Hum. Genet*. 12: 220–224.

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.

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.

Kimura M., 1985 The role of compensatory neutral mutations in molecular evolution. *J. Genet*. 64: 7.

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.

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.

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.

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.

Shannon C. E., 1948 A Mathematical Theory of Communication. *Bell System Technical Journal* 27: 379–423.

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.

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.

Williams G. C., 1957 Pleiotropy, Natural Selection, and the Evolution of Senescence. *Evolution* 11: 398–411.

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, pp. 97–166 in *Evolving Genes and Proteins*.

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.

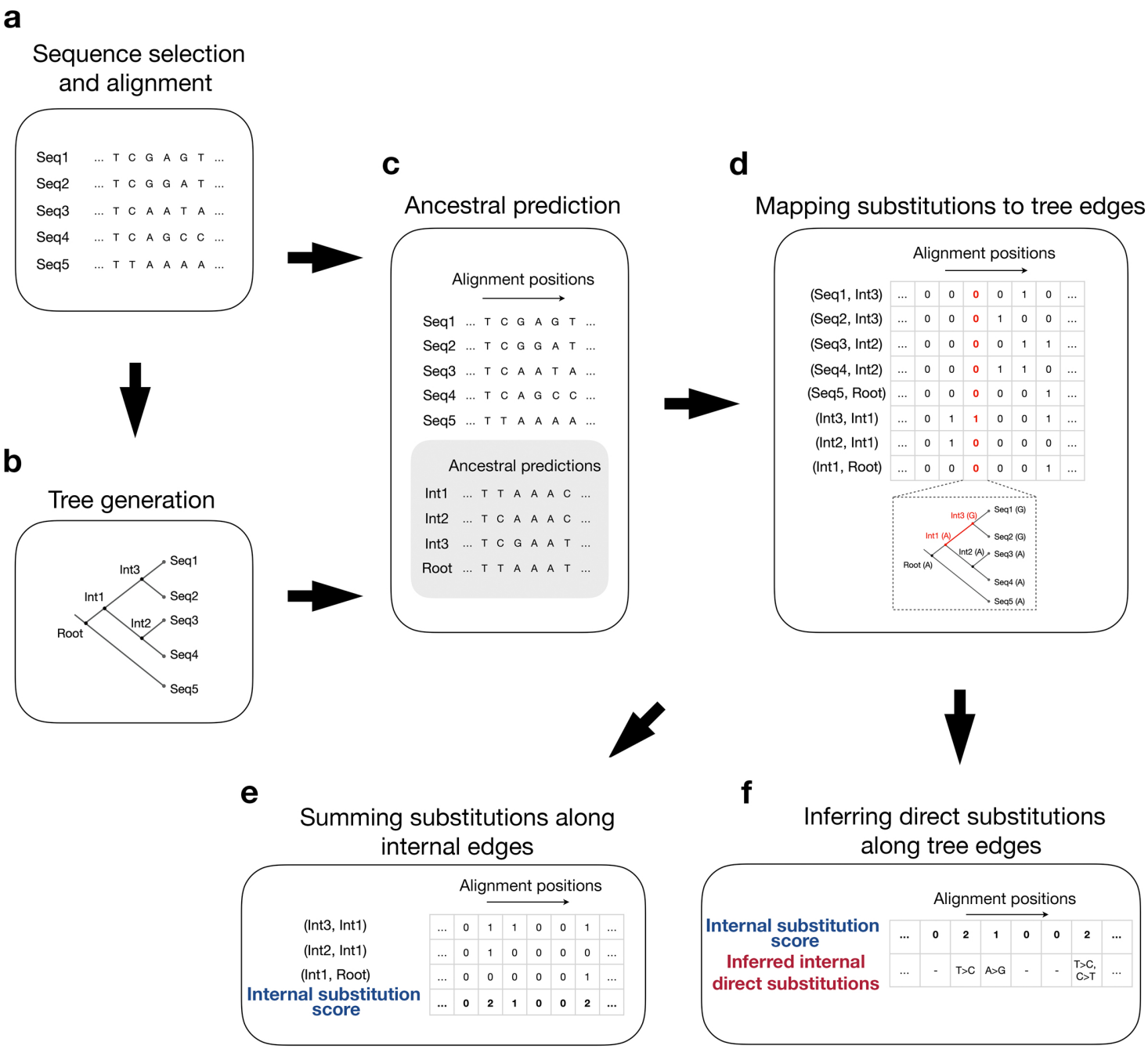

a
Sequence selection and alignment
Seq1 ... T C G A G T ...
Seq2 ... T C G G A T ...
Seq3 ... T C A A T A ...
Seq4 ... T C A G C C ...
Seq5 ... T T A A A A ...
b
Tree generation
Root
Int1
Int2
Int3
Seq1
Seq2
Seq3
Seq4
Seq5
c
Ancestral prediction
Alignment positions
Seq1 ... T C G A G T ...
Seq2 ... T C G G A T ...
Seq3 ... T C A A T A ...
Seq4 ... T C A G C C ...
Seq5 ... T T A A A A ...
Ancestral predictions
Int1 ... T T A A A C ...
Int2 ... T C A A A C ...
Int3 ... T C G A A T ...
Root ... T T A A A T ...
d
Mapping substitutions to tree edges
Alignment positions
(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 (G)
Int1 (A)
Int2 (A)
Root (A)
Seq1 (G)
Seq2 (G)
Seq3 (A)
Seq4 (A)
Seq5 (A)
e
Summing substitutions along internal edges
Alignment positions
(Int3, Int1) ... 0 1 1 0 0 1 ...
(Int2, Int1) ... 0 1 0 0 0 0 ...
(Int1, Root) ... 0 0 0 0 0 1 ...
Internal substitution score ... 0 2 1 0 0 2 ...
f
Inferring direct substitutions along tree edges
Alignment positions
Internal substitution score ... 0 2 1 0 0 2 ...
Inferred internal direct substitutions ... - T>C A>G - - T>C, C>T ...

Figure 1

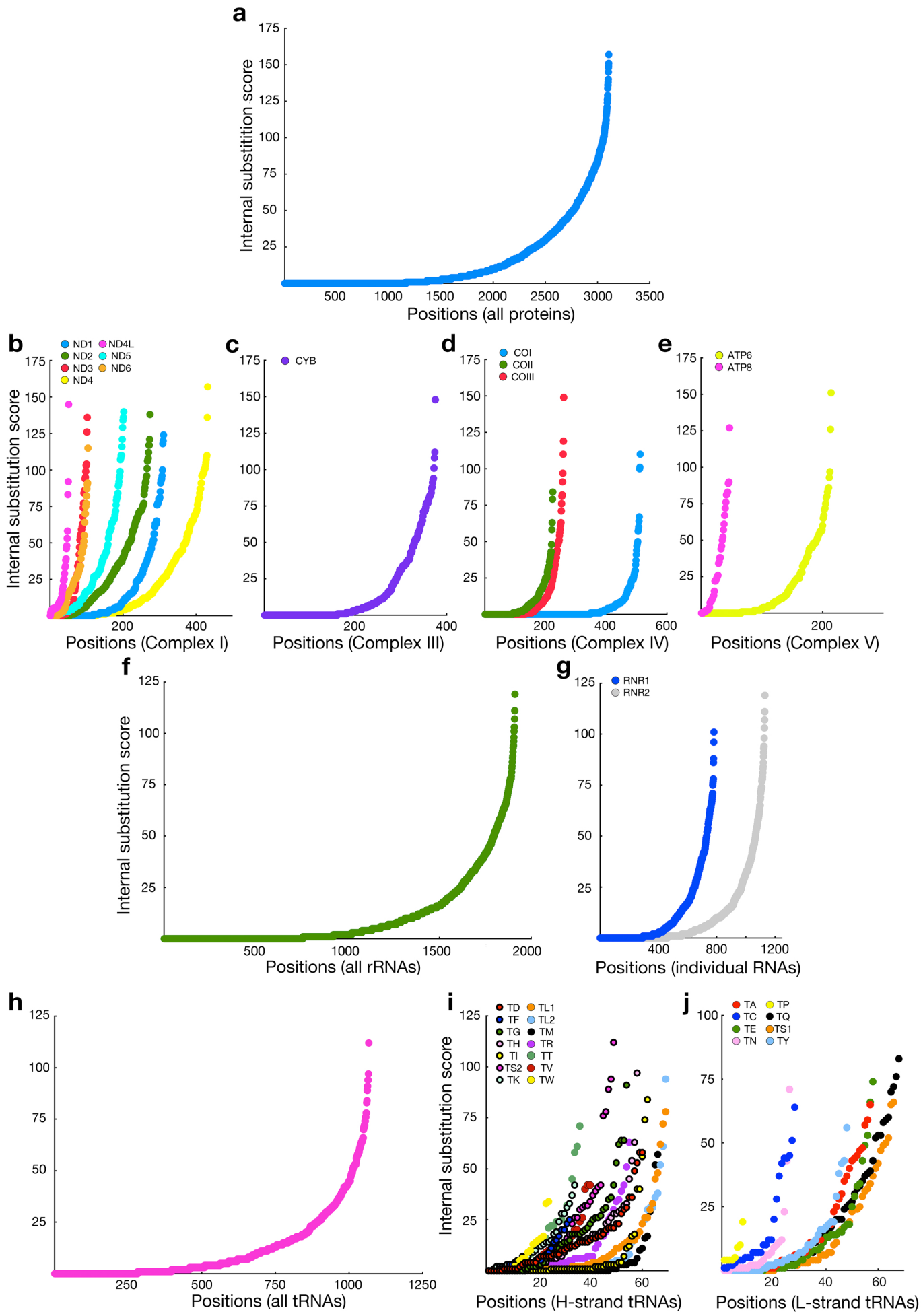
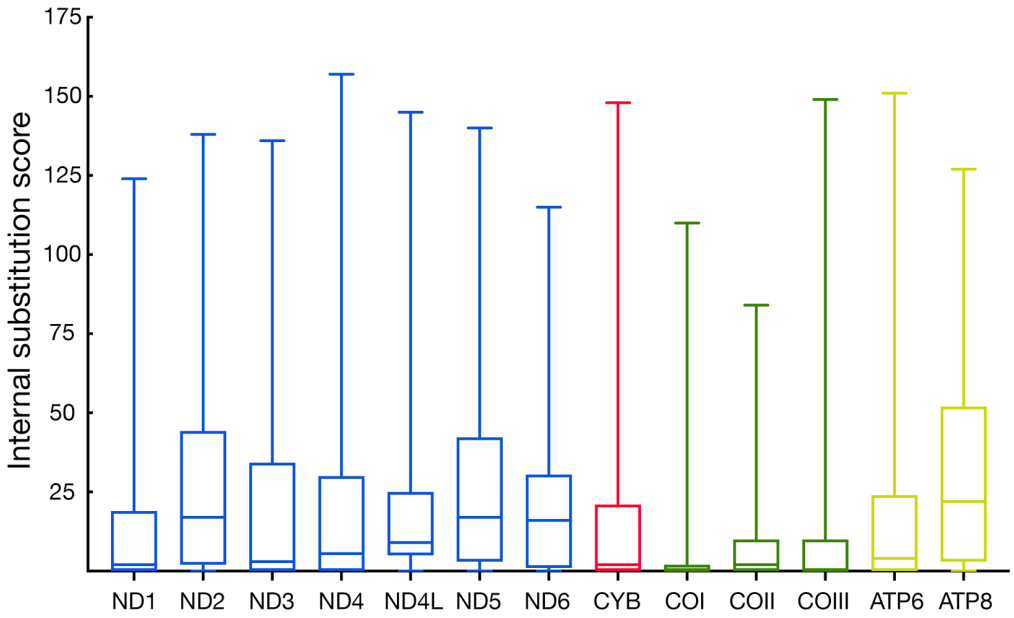

a
Internal substitition score
Positions (all proteins)
b
ND1
ND2
ND3
ND4
ND4L
ND5
ND6
Internal substitution score
Positions (Complex I)
c
CYB
Positions (Complex III)
d
COI
COII
COIII
Positions (Complex IV)
e
ATP6
ATP8
Positions (Complex V)
f
Internal substitution score
Positions (all rRNAs)
g
RNR1
RNR2
Positions (individual RNAs)
h
Positions (all tRNAs)
i
TD
TF
TG
TH
TI
TS2
TK
TL1
TL2
TM
TR
TT
TV
TW
Internal substitution score
Positions (H-strand tRNAs)
j
TA
TC
TE
TN
TP
TQ
TS1
TY
Positions (L-strand tRNAs)

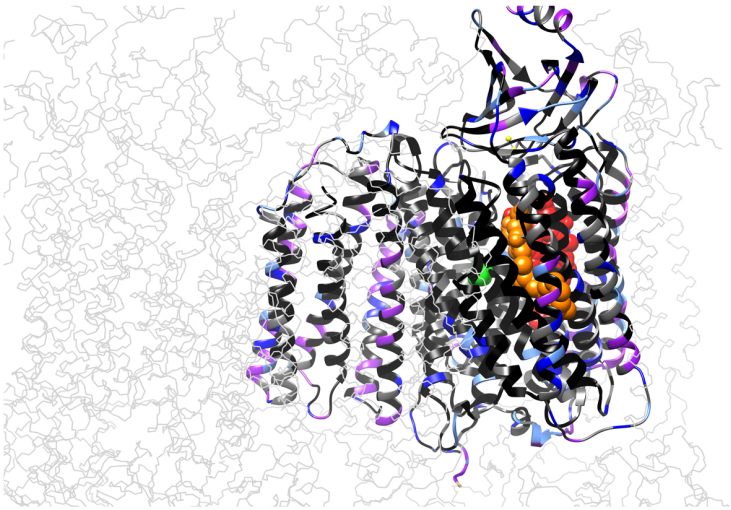
Figure 2

a



b

side view



c

top view

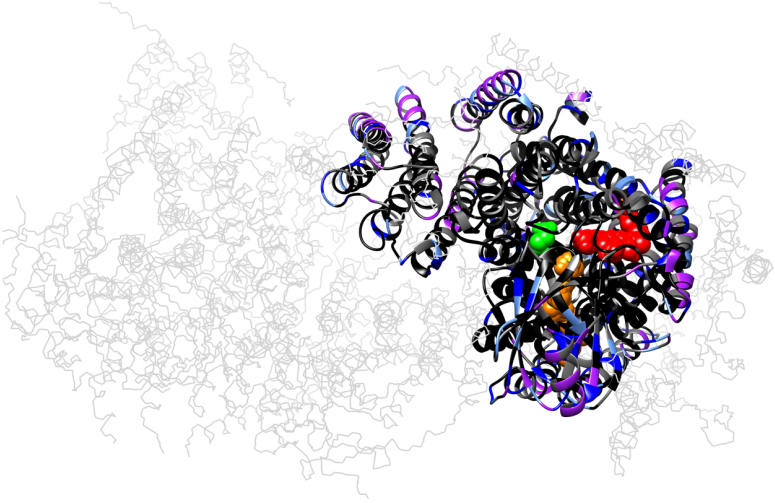


Figure 3

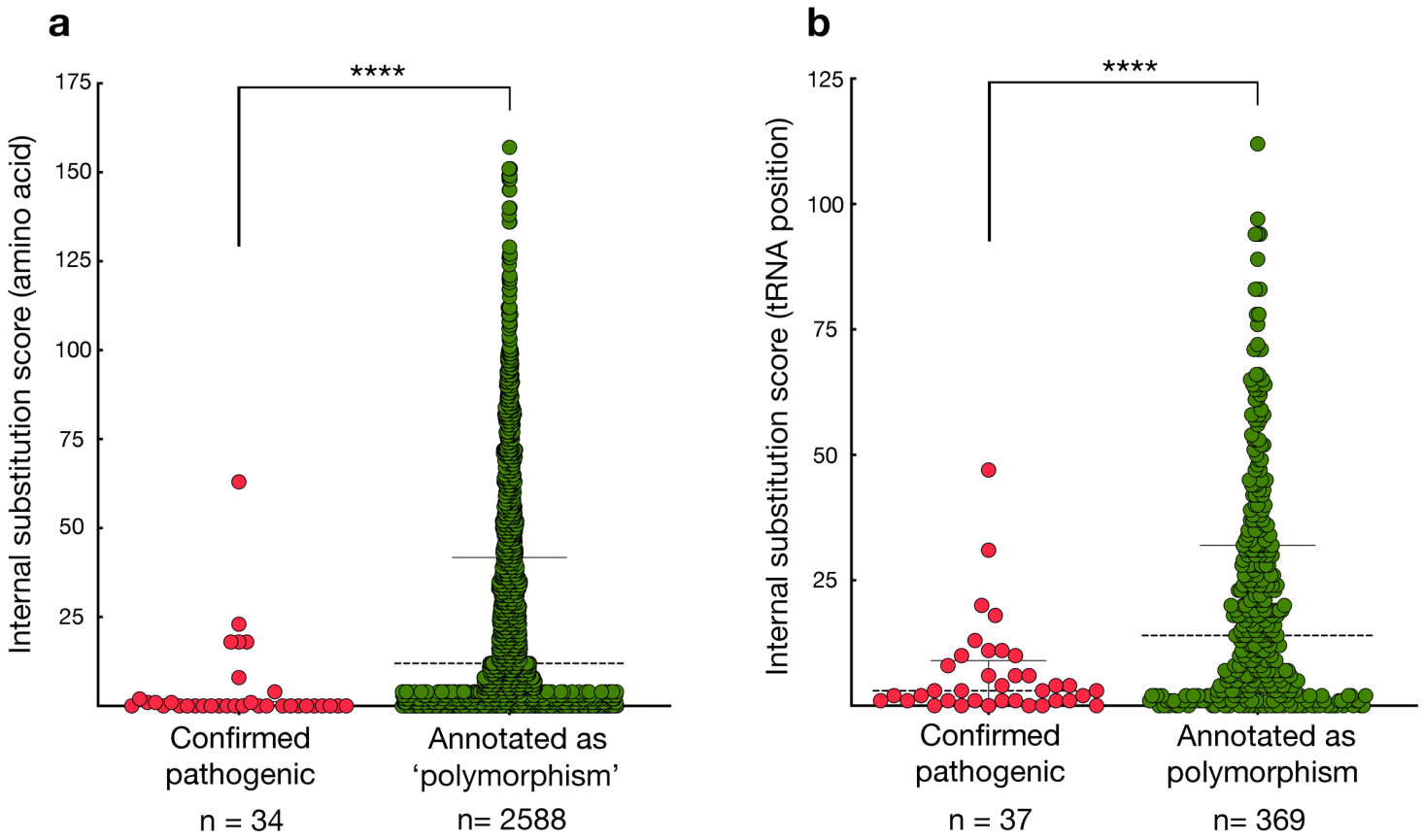

a
****
175
150
125
100
75
50
25
Internal substitution score (amino acid)
Confirmed pathogenic
n = 34
Annotated as 'polymorphism'
n= 2588
b
****
125
100
75
50
25
Internal substitution score (tRNA position)
Confirmed pathogenic
n = 37
Annotated as polymorphism
n= 369

Figure 4

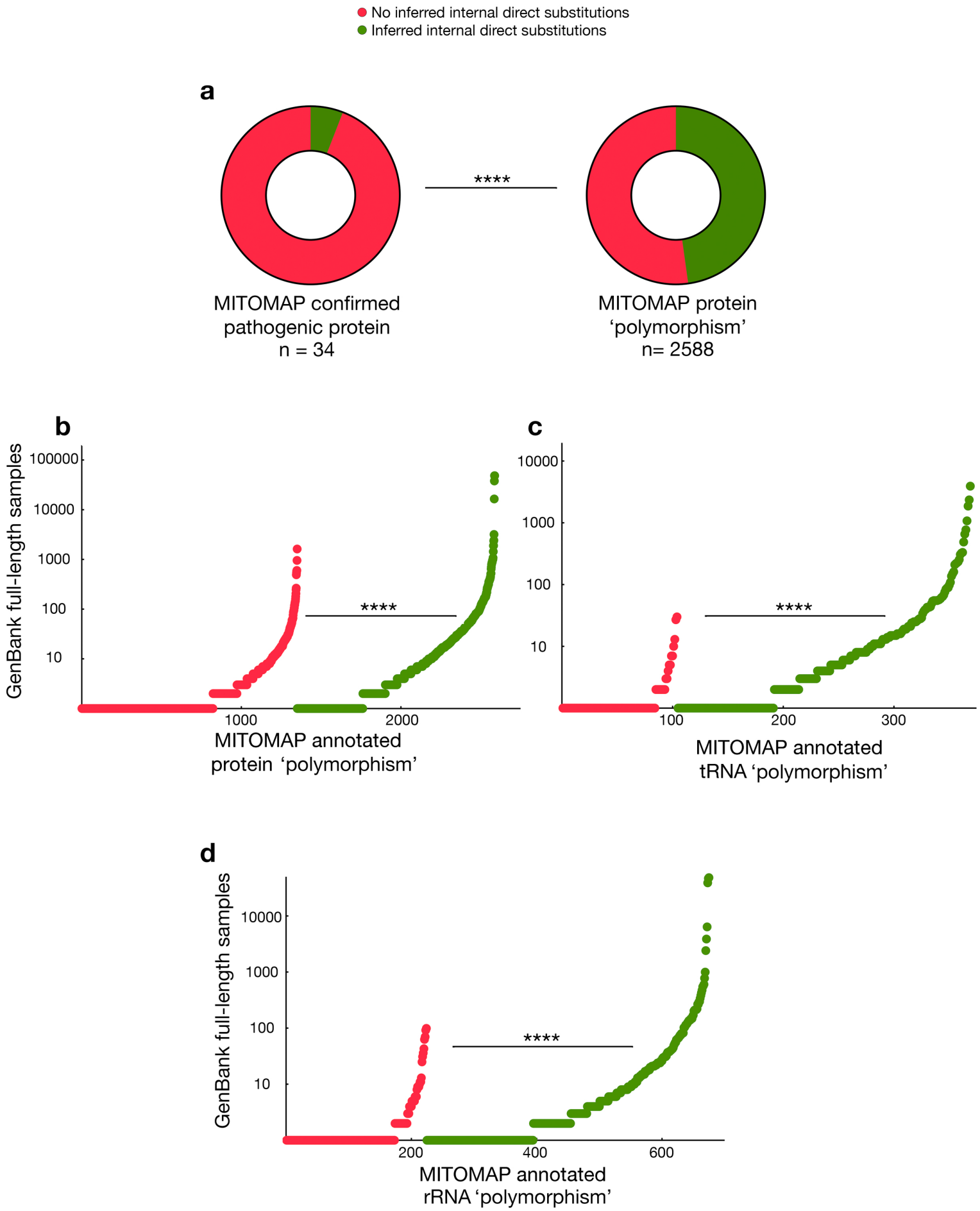

No inferred internal direct substitutions
Inferred internal direct substitutions
a
****
MITOMAP confirmed pathogenic protein n = 34
MITOMAP protein 'polymorphism' n= 2588
b
GenBank full-length samples
100000
10000
1000
100
10
****
1000
2000
MITOMAP annotated protein 'polymorphism'
c
10000
1000
100
10
****
100
200
300
MITOMAP annotated tRNA 'polymorphism'
d
GenBank full-length samples
10000
1000
100
10
****
200
400
600
MITOMAP annotated rRNA 'polymorphism'

Figure 5

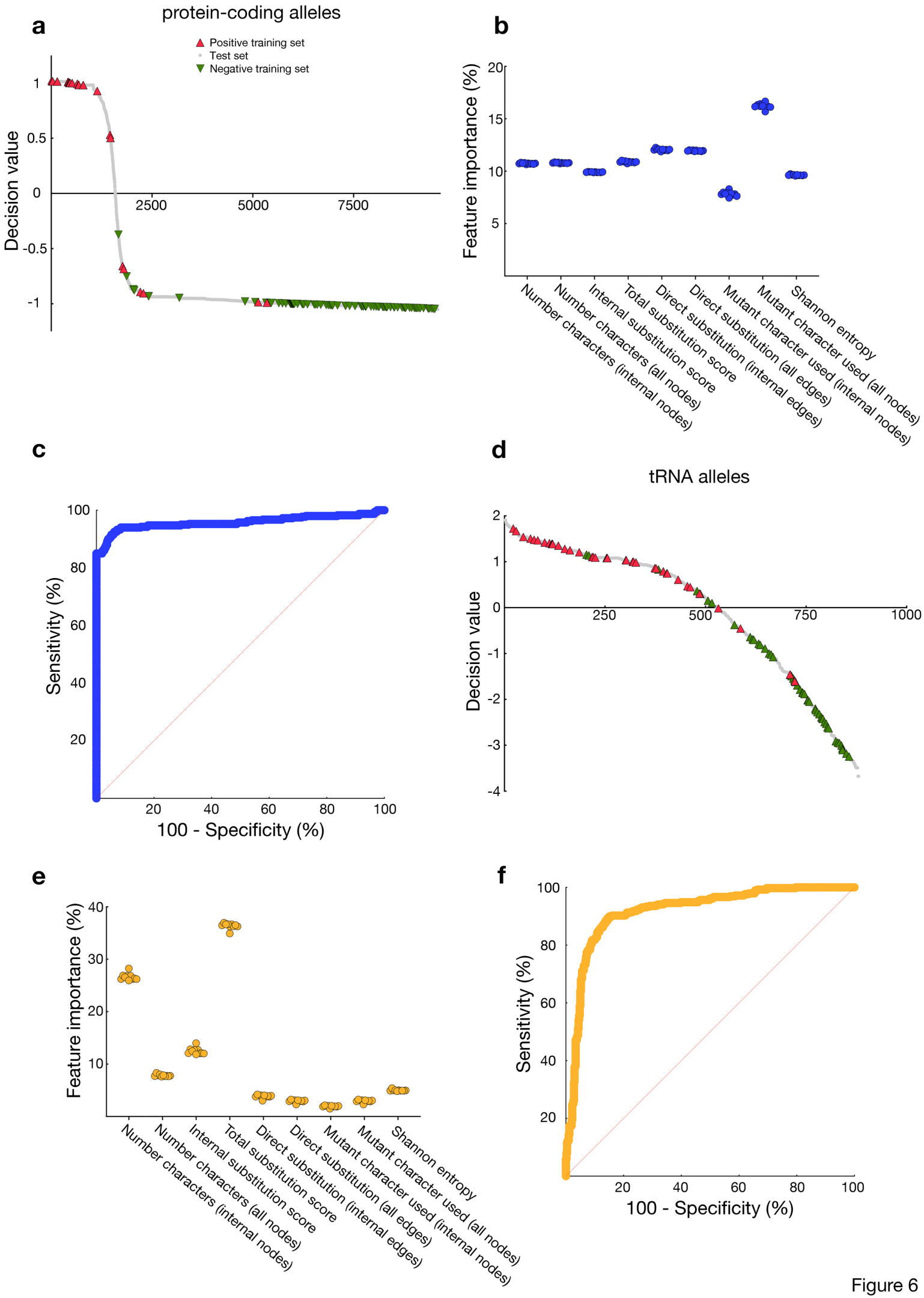

a
protein-coding alleles
Positive training set
Test set
Negative training set
Decision value
1
0.5
0
-0.5
-1
2500
5000
7500
b
Feature importance (%)
20
15
10
5
Number characters (internal nodes)
Number characters (all nodes)
Internal substitution score
Total substitution score
Direct substitution (internal edges)
Direct substitution (all edges)
Mutant character used (internal nodes)
Mutant character used (all nodes)
Shannon entropy
c
Sensitivity (%)
100
80
60
40
20
100 - Specificity (%)
d
tRNA alleles
Decision value
2
1
0
-1
-2
-3
-4
250
500
750
1000
e
Feature importance (%)
40
30
20
10
Number characters (internal nodes)
Number characters (all nodes)
Internal substitution score
Total substitution score
Direct substitution (internal edges)
Direct substitution (all edges)
Mutant character used (internal nodes)
Mutant character used (all nodes)
Shannon entropy
f
Sensitivity (%)
100
80
60
40
20
100 - Specificity (%)

Figure 6

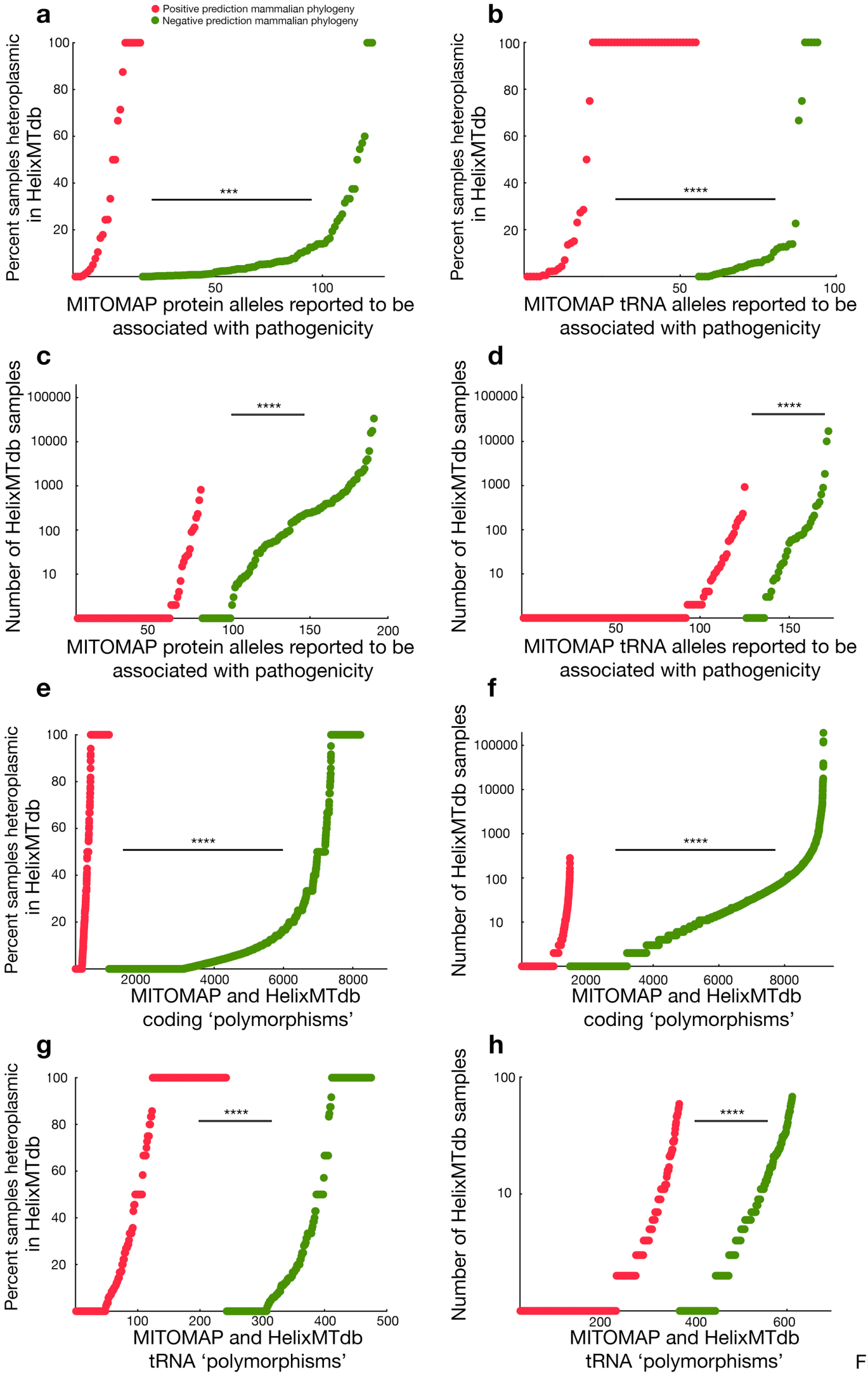

a
Positive prediction mammalian phylogeny
Negative prediction mammalian phylogeny
Percent samples heteroplasmic in HelixMTdb
***
MITOMAP protein alleles reported to be associated with pathogenicity
b
Percent samples heteroplasmic in HelixMTdb
****
MITOMAP tRNA alleles reported to be associated with pathogenicity
c
Number of HelixMTdb samples
****
MITOMAP protein alleles reported to be associated with pathogenicity
d
Number of HelixMTdb samples
****
MITOMAP tRNA alleles reported to be associated with pathogenicity
e
Percent samples heteroplasmic in HelixMTdb
****
MITOMAP and HelixMTdb coding 'polymorphisms'
f
Number of HelixMTdb samples
****
MITOMAP and HelixMTdb coding 'polymorphisms'
g
Percent samples heteroplasmic in HelixMTdb
****
MITOMAP and HelixMTdb tRNA 'polymorphisms'
h
Number of HelixMTdb samples
****
MITOMAP and HelixMTdb tRNA 'polymorphisms'

Figure 7

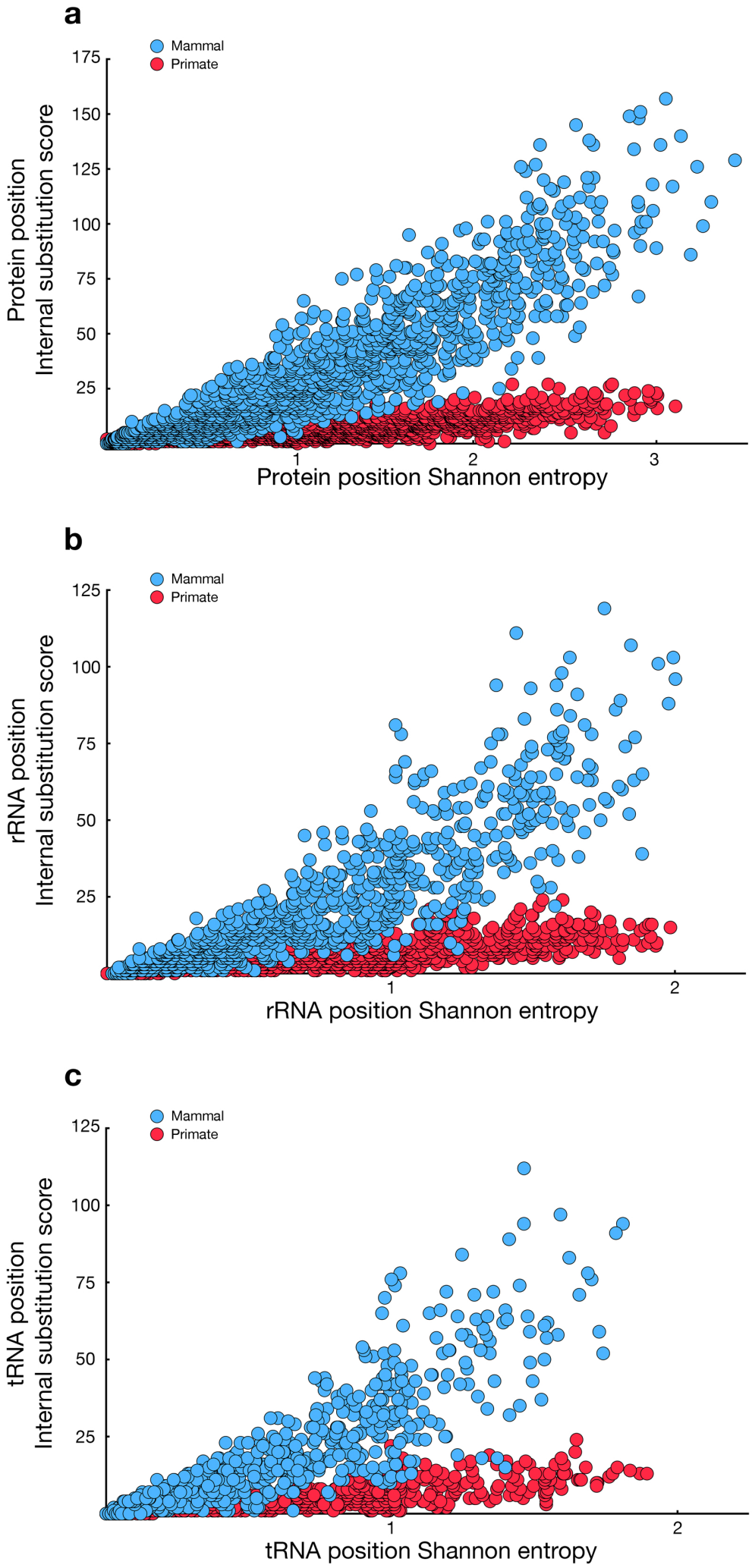

a
Mammal
Primate
Protein position
Internal substitution score
175
150
125
100
75
50
25
1
2
3
Protein position Shannon entropy
b
Mammal
Primate
rRNA position
Internal substitution score
125
100
75
50
25
1
2
rRNA position Shannon entropy
c
Mammal
Primate
tRNA position
Internal substitution score
125
100
75
50
25
1
2
tRNA position Shannon entropy

Figure S1

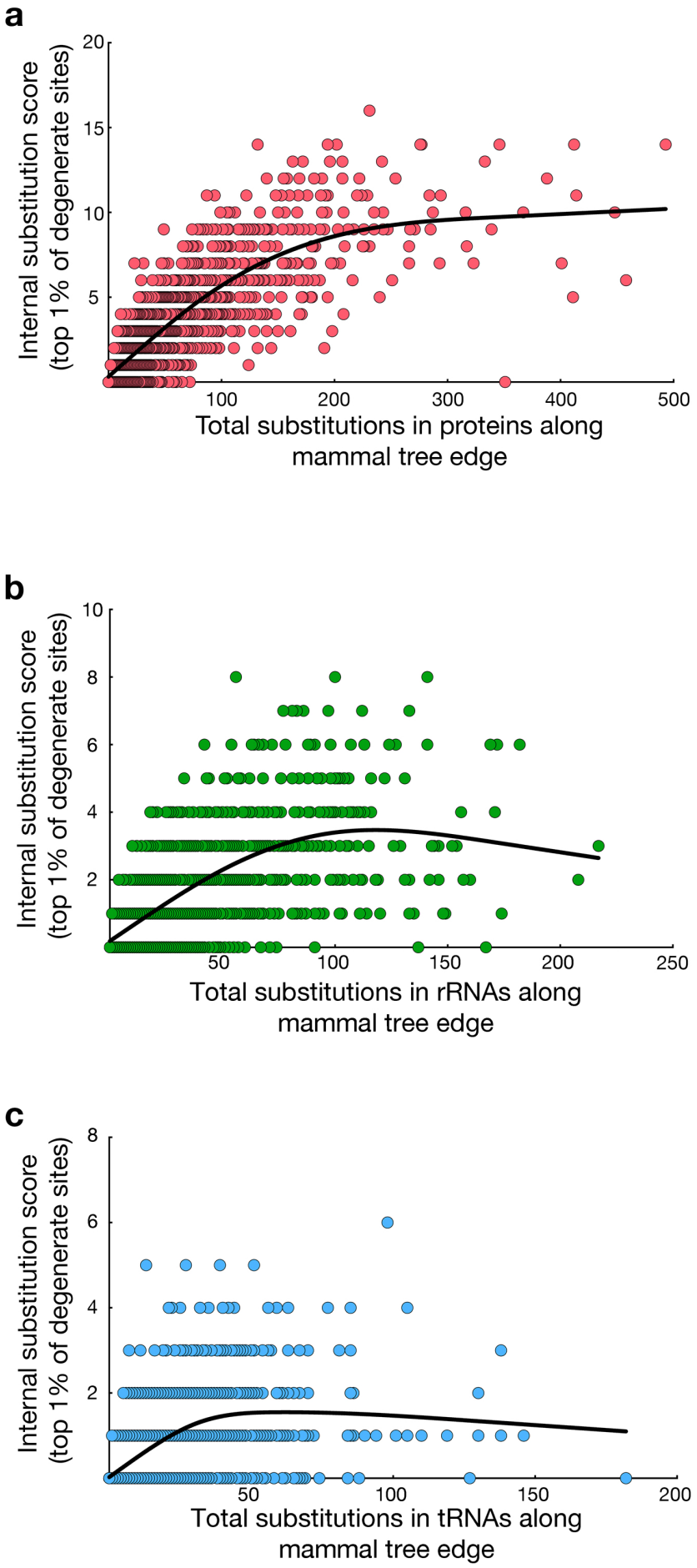

a
Internal substitution score
(top 1% of degenerate sites)
20
15
10
5
100
200
300
400
500
Total substitutions in proteins along
mammal tree edge
b
Internal substitution score
(top 1% of degenerate sites)
10
8
6
4
2
50
100
150
200
250
Total substitutions in rRNAs along
mammal tree edge
c
Internal substitution score
(top 1% of degenerate sites)
8
6
4
2
50
100
150
200
Total substitutions in tRNAs along
mammal tree edge

Figure S2