© 2021. Published by The Company of Biologists Ltd | Biology Open (2021) 10, bio059072. doi:10.1242/bio.059072

RESEARCH ARTICLE

The population frequency of human mitochondrial DNA variants is highly dependent upon mutational bias

Cory D. Dunn

ABSTRACT

Next-generation sequencing can quickly reveal genetic variation potentially linked to heritable disease. As databases encompassing human variation continue to expand, rare variants have been of high interest, since the frequency of a variant is expected to be low if the genetic change leads to a loss of fitness or fecundity. However, the use of variant frequency when seeking genomic changes linked to disease remains very challenging. Here, I explored the role of selection in controlling human variant frequency using the HelixMT database, which encompasses hundreds of thousands of mitochondrial DNA (mtDNA) samples. I found that a substantial number of synonymous substitutions, which have no effect on protein sequence, were never encountered in this large study, while many other synonymous changes are found at very low frequencies. Further analyses of human and mammalian mtDNA datasets indicate that the population frequency of synonymous variants is predominantly determined by mutational biases rather than by strong selection acting upon nucleotide choice. My work has important implications that extend to the interpretation of variant frequency for non-synonymous substitutions.

KEY WORDS: Genomic variation, Mitochondrial DNA, Mutational bias, Pathogenicity prediction, Population frequency

INTRODUCTION

In this era of genomic medicine, next-generation sequence data obtained from patients, families, and populations are used to reveal and predict which genes and variants may be linked to disease (Claussnitzer et al., 2020; Shendure et al., 2019). Researchers have often embraced rare variants when seeking genomic changes in protein-coding sequences that may be pathogenic, as a reduction in variant frequency is an expected outcome of selection (Bomba et al., 2017; Gibson, 2012; Sazonovs and Barrett, 2018; Zuk et al., 2014). However, while triage of rare variants has led to some success in illuminating genes linked to heritable disease (Fuchsberger et al., 2016; Genovese et al., 2016; Lencz et al., 2021; Luo et al., 2017), the interpretation and utilization of rare genomic changes remains very challenging (Macklin et al., 2018; Manrai et al., 2016; Uricchio et al., 2016).

Human mitochondrial DNA (mtDNA) encodes proteins and RNAs required for the process of oxidative phosphorylation, and mitochondrial mutations are linked to a number of metabolic diseases (Gorman et al., 2016; Thompson et al., 2020). Recently, the mtDNAs of nearly 200,000 individuals were sequenced in order to produce the HelixMT database (HelixMTdb), a large catalog of human mtDNA variation (Bolze et al., 2019 preprint). Here, I found that many synonymous nucleotide substitutions were never detected within this quite substantial survey of human mtDNA. Subsequent study of more than one thousand mammalian mtDNAs suggested that selection on synonymous base substitution in mitochondrial protein-coding genes is minimal and unlikely to explain the absence or rarity of many synonymous changes within the human population. Rather, the mutational propensities of mtDNA (Kumar, 1996; Reyes et al., 1998) are more likely to have a dominant influence upon variant frequency. My findings have general implications for the interpretation of variant frequencies when studying heritable disease.

RESULTS

During an exploration of selective pressures that may act upon mitochondria-encoded polypeptides, I simulated every possible nucleotide substitution from the human reference mtDNA sequence within all protein-coding sequences, then cross-referenced these nucleotide and potential amino acid changes with the nucleotide substitutions tabulated in HelixMTdb. Consistent with selection against the vast majority of amino acid changes, non-synonymous substitutions were depleted to a far greater extent than synonymous substitutions when considering either the number of samples harboring variants of each type (Fig. 1A) or whether the nucleotide substitution was encountered at all during compilation of the HelixMTdb (Fig. 1B).

Since synonymous changes would not be expected to change the structure or function of mitochondria-encoded proteins, the number of changes (2925, or ∼35% of potential synonymous substitutions) for which a synonymous substitution was never encountered (abbreviated here as 'SSNEs') during the HelixMTdb investigation was considered noteworthy. While HelixMTdb contained samples from almost all human haplogroups (Bolze et al., 2019 preprint), and while closely related individuals were removed from the study, more than 90% of samples were classified within the 'N' macro-haplogroup, a lineage associated with the exit of modern humans from Africa (Ingman et al., 2000; Maca-Meyer et al., 2001). Consequently, it was possible that SSNE abundance was linked to limited sample diversity. Therefore, in an attempt to assess HelixMTdb sampling biases, as well as to explore this dataset more deeply, I examined in greater detail the presence or absence of HelixMTdb samples harboring each synonymous change to a third codon position (abbreviated as a 'P3'). When considering those amino acids associated only with twofold P3 degeneracy, or the ability to accept only two different bases at P3 without changing the protein sequence, samples diverging from the human reference sequence were present in HelixMTdb for nearly all (>97%)

Institute of Biotechnology, University of Helsinki, Helsinki 00014, Finland.

*Author for correspondence (cory.dunn@helsinki.fi)

C.D.D., 0000-0003-2393-5944

Received 30 September 2021; Accepted 7 October 2021

Biology Open (2021) 10, bio059072. doi:10.1242/bio.059072

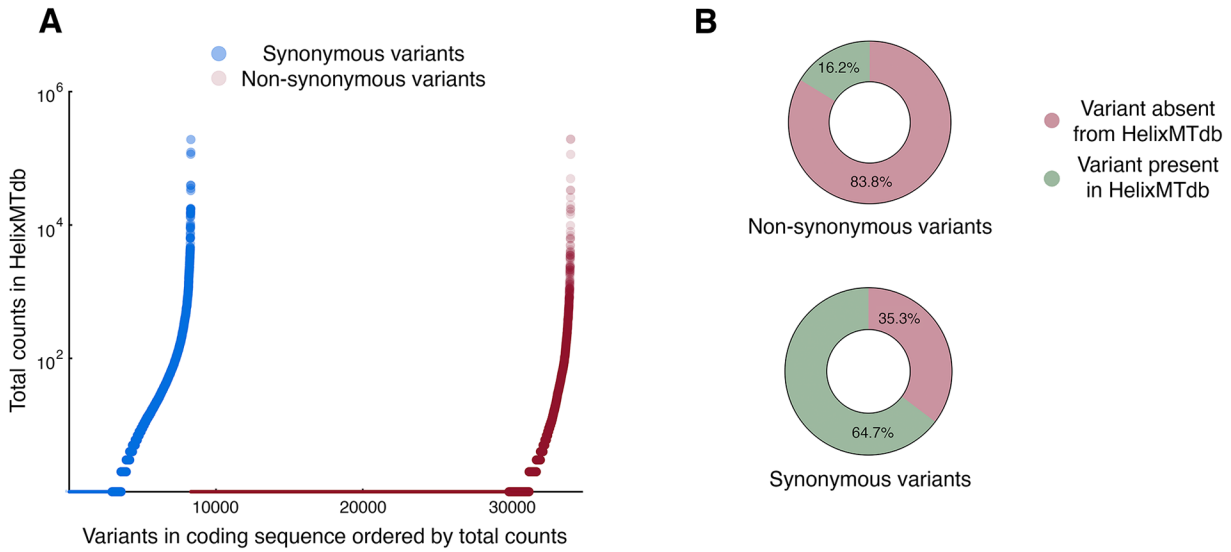

Fig. 1. Many synonymous variants are never encountered within the HelixMT database. (A) Non-synonymous and synonymous variants differ in their population frequency. The sample count for variants within coding regions that are reachable by a single substitution from the human mtDNA reference sequence were obtained from HelixMTdb and plotted (Kolmogorov–Smirnov approximate *P*-value, <0.0001). Non-synonymous and synonymous substitutions never encountered are reflected as smaller dots along the x-axis at the zero position on the y-axis. (B) While the great majority of non-synonymous substitutions were never encountered during the HelixMTdb study, a substantial fraction of synonymous substitutions are also apparently lacking from the human population.

analyzed positions (Fig. 2A). This result indicates that the HelixMTdb does indeed cover a substantial amount of human mtDNA sequence diversity. Furthermore, our analysis demonstrates that SSNEs are unlikely to be associated with twofold degenerate sites and that codon choice at twofold degenerate sites governed by transitions (a purine–purine change or a pyrimidine–pyrimidine change) is unlikely to be under strong selection in humans. Since transitions at twofold degenerate sites cause changes in local GC content, as well as slight alterations in global nucleotide content, our results also argue against substantial selection upon human mtDNA that might be based upon these factors.

Next, I examined variation at amino acids associated only with fourfold P3 degeneracy, or the ability to accept all nucleotides at P3 without altering the protein sequence. Nearly all analyzed fourfold degenerate positions (99%) harbored at least one sample with a base differing from that found in the human reference (Fig. 2B), again suggesting that HelixMTdb encompasses a sizable amount of human mtDNA diversity. However, I encountered apparent limitations on base variability at some of these fourfold degenerate P3s, as all four base possibilities could be identified within HelixMTdb at only 10.4% of these mtDNA locations.

I then tested whether SSNEs might be predominantly associated with specific amino acids. I found that for the amino acids arginine, threonine, alanine, valine, serine, glycine, proline, and leucine, a more substantial fraction of synonymous changes were never seen in the HelixMTdb relative to the other amino acids (Fig. 2C). All of these amino acids can be associated with vertebrate mtDNA codons of fourfold degeneracy.

While twofold degenerate P3s allow only transitions without altering the amino acid, fourfold degenerate P3s also permit transversions (a purine-pyrimidine change, or vice versa) that leave the protein code unchanged. Since amino acids with fourfold degenerate P3s were characterized by a higher number of SSNEs, I tested whether SSNEs might be more closely associated with transversions or transitions. I found that nearly all SSNEs assigned to these eight amino acids (>96%) were linked to a potential transversion (Fig. 2D). Expanding my analysis to synonymous substitutions encountered at least once in HelixMTdb, the population frequency of a variant was also clearly linked to whether the nucleotide change at fourfold degenerate P3s was a transition or a transversion (Fig. 2E).

Two conspicuous and non-exclusive hypotheses exist regarding the notable enrichment of transversions among SSNEs at fourfold degenerate P3s. First, there may be unanticipated, yet substantial selection that acts upon P3s and leads to depletion of even synonymous transversions from the human population. Second, mutational biases related to mtDNA replication and maintenance may make transversions at degenerate P3s far less likely than transitions (Aquadro and Greenberg, 1983; Belle et al., 2005; Brown and Simpson, 1982; Kennedy et al., 2013; Kumar, 1996; Tamura and Nei, 1993; Vermulst et al., 2007; Wakeley, 1996; Zaidi et al., 2019). To address the first possibility, I further examined the extent of selection on P3s among mammals by examining the nucleotide frequencies at approximately 5 million P3s across the coding sequences of 1317 mammalian mtDNAs. Here, I also took into account the two different mitochondrial tRNAs recognizing leucine codons and the two mitochondrial tRNAs recognizing serine codons. As encountered in previous studies (Kumar, 1996; Reyes et al., 1998), guanine was depleted from mitochondrial P3s (the vast majority of which are encoded by the L-strand) for which the presence of any purine does not lead to an amino acid change (Fig. 3A), while adenine dominated at those positions. Cytosine and thymine were both well-represented at P3s for which any pyrimidine is permitted without altering the encoded amino acid. However, even considering the relative depletion of guanine from all fourfold degenerate P3s and twofold degenerate purine P3s, guanine was nonetheless detected at thousands of P3 positions (Fig. 3B). Consequently, nucleotide frequencies at P3s appear unlikely to reflect generalized codon limitations inherent to the process of mitochondrial translation.

While degeneracy at the third codon position is a general feature of mtDNA-encoded amino acids, these results are not necessarily

Biology Open (2021) 10, bio059072. doi:10.1242/bio.059072

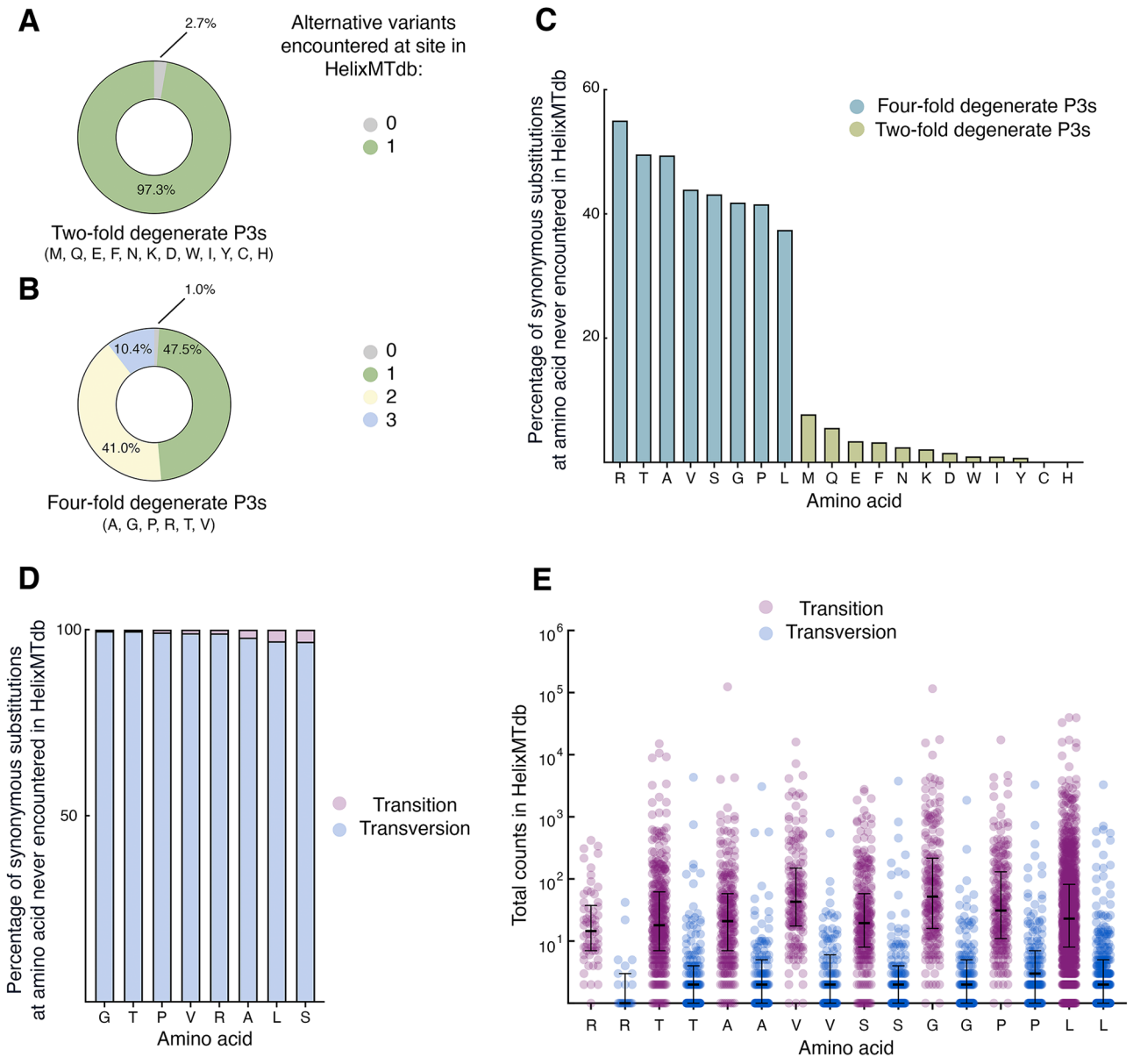

Fig. 2. Transversion substitutions from the reference sequence are heavily depleted at fourfold degenerate third-codon positions of human mitochondrial coding sequences. (A) Nearly all synonymous changes from the human reference sequence can be identified within HelixMTdb at amino acid positions characterized by twofold degenerate P3s. (B) Nearly all amino acid positions characterized by fourfold degeneracy at P3 could be associated with at least one base substitution in the HelixMTdb dataset. (C) Synonymous mutations never encountered in the HelixMTdb study are most abundant at amino acids for which at least one codon is fourfold degenerate at P3. (D) The vast majority of synonymous substitutions never encountered by the HelixMTdb study are transversions. (E) For those substitutions that are encountered in HelixMTdb, population prevalence is linked to substitution type. For each transition or transversion found in the HelixMTdb at an amino acid for which at least one codon is fourfold degenerate at P3, the HelixMTdb population count is plotted. Bar and error bars represent median of population count and interquartile range, respectively. For all comparisons of variant frequencies for a given amino acid (transition versus transversion from reference), Kolmogorov–Smirnov approximate *P*-values are <0.0001.

informative about the possibility that strong selection acts upon *specific* P3s encoded by the mitochondrial genome. To further explore the extent to which individual P3s might be under selection, I focused my attention upon codons for which the first and second positions, and therefore the encoded amino acids, are 100% identical in an alignment consisting of 1251 mammals and an outlier reptile sequence (*Iguana iguana*) used to root an inferred phylogenetic tree. Leucine codons could not be included in this analysis, as degeneracy at the first codon position always led to substitution between codons recognized by the L1 and L2 tRNAs at mammalian protein alignment sites harboring only leucine. I found that 560/561 (99.8%) of the resulting set of P3s ('I-P3s', indicating identity of codon positions one and two throughout the alignment) can be inhabited by any nucleotide permitting synonymous substitution (Fig. 3C). Only the I-P3 of the codon annotated in humans as encoding the COX3 starting methionine appears to be totally constrained with respect to nucleotide choice, as this P3 is always occupied by guanine in mammals. Interestingly, COX3 is reported to be unique among mitochondrial polypeptides for its lack of an amino-terminal formyl-methionine (Walker et al., 2009), although whether there is a mechanistic relationship between these two observations remains to be determined. Nearly complete

Biology Open (2021) 10, bio059072. doi:10.1242/bio.059072

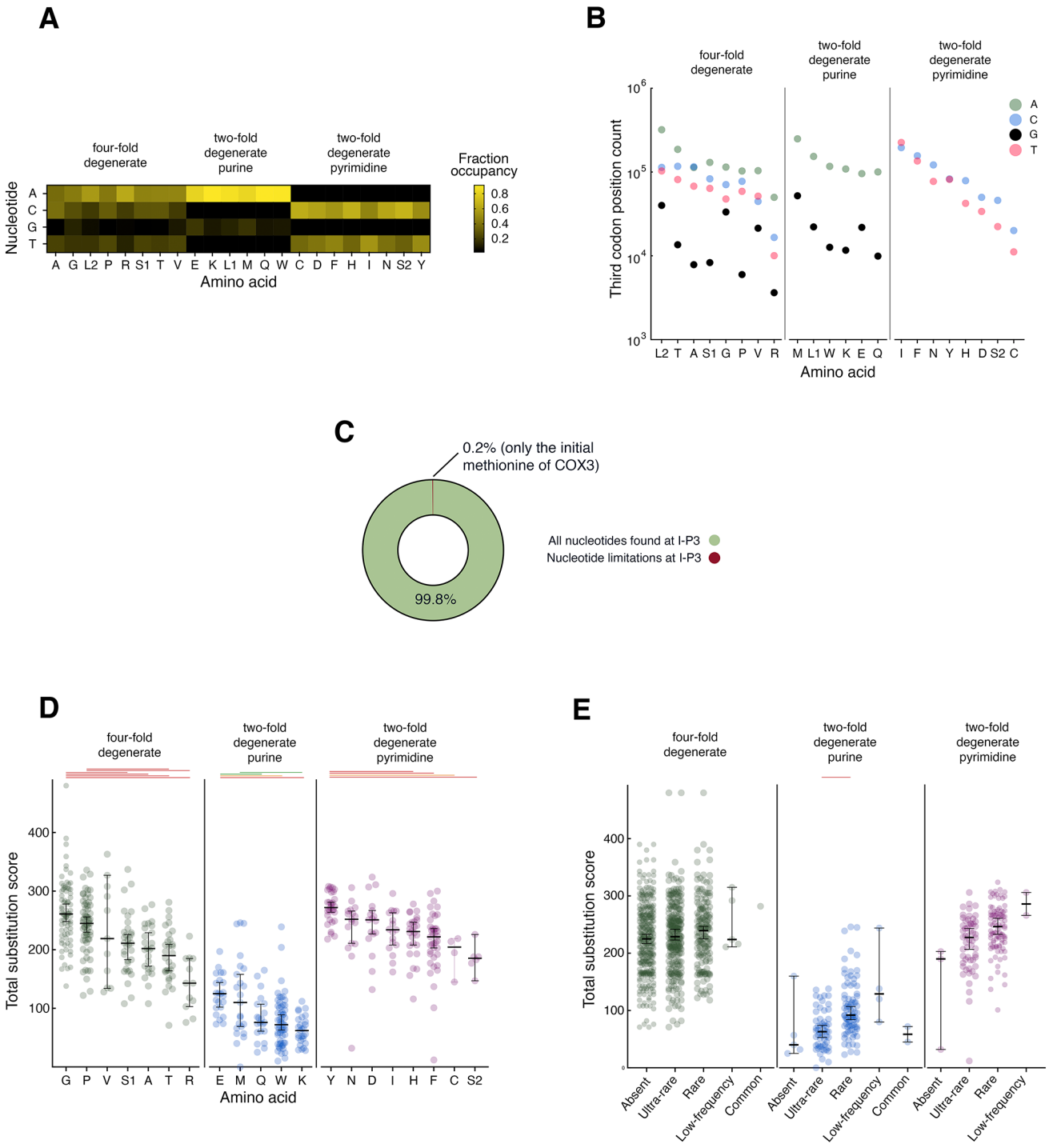

Fig. 3. Abundant substitution and degeneracy across the third codon positions of mammals. (A) Base occupancy is not equally distributed among nucleotides at degenerate P3s ($n \geq 33{,}073$ instances of each amino acid are analyzed across the set of input mammalian mtDNAs). (B) Guanine is not excluded from recognition by any tRNA presumed to allow silent purine substitution at P3. (C) Nearly all I-P3s can accept all of the possible synonymous substitutions available for a particular amino acid. (D) Substitution is common at nearly all mammalian I-P3 positions, although the TSS distributions of I-P3s associated with amino acids can differ within each degeneracy class (fourfold degenerate, twofold degenerate purine, or twofold degenerate pyrimidine). Kolmogorov–Smirnov approximate *P*-values corrected for multiple comparisons are shown (red, ≤0.001; orange, ≤0.01; green, ≤0.05; no line,>0.05). Bar and error bars represent median with 95% confidence interval. (E) Population frequency of synonymous variants is unlinked to TSS at fourfold degenerate I-P3s. Statistical significance is represented as in D, and the bar and error bars represent median and 95% confidence intervals.

degeneracy among mammalian I-P3s again argues against strong selection upon synonymous nucleotide substitution within mammals based upon codon preference and the thermodynamics of base pairing, while also arguing against selection due to the presence of conserved protein binding sites.

I then calculated for each I-P3 the total substitution score (TSS), a sum of substitutions occurring at a specific site throughout an inferred mammalian phylogenetic tree (Akpınar et al., 2021 preprint). I found, with few exceptions, that nearly all fourfold degenerate, twofold degenerate pyrimidine, and twofold degenerate

Biology Open (2021) 10, bio059072. doi:10.1242/bio.059072

purine I-P3s have been subject to base substitution tens, or even hundreds of times, during approximately 200 million years of mammal evolution (Fig. 3D), a result quite consistent with minimal selection upon nucleotide choice at mitochondrial P3s. However, here I did note statistically significant divergence between TSS distributions at I-P3s when comparing amino acids within a given degeneracy class. These results suggest, when also considering the findings described above, that while synonymous base substitutions are not under strong selective pressure, weak selection may have helped to shape some mitochondrial third codon positions throughout the course of mammalian evolution.

Next, I asked whether a low frequency of human variants at fourfold degenerate I-P3s would correspond with lower mammal-derived TSS values of corresponding positions, a potential indicator of selection extending across mammals to humans. Here, I placed variation occurring at I-P3s into the classes 'absent' (zero counts among 195983 HelixMTdb samples), 'ultra-rare' (variant frequency <0.01%), 'rare' (variant frequency <1% and ≥0.01%), 'low-frequency' (variant frequency ≥1% and <than 5%) or 'common' (variant frequency ≥to 5%). However, I detected no significant relationship between TSS and variant frequency for fourfold degenerate and twofold degenerate pyrimidine I-P3s (Fig. 3E), suggesting that the prominent SSNE abundance at fourfold degenerate P3s is unlikely to be due to selection. A statistically significant link after correction for multiple testing was only observed when comparing the TSS distribution between ultra-rare and rare variants at analyzed I-P3s harboring purines, providing limited evidence of mammal-wide selection that might be linked to the frequency of human mtDNA variation.

Finally, I explored how the tabulation of heteroplasmic samples within the HelixMTdb might be informative regarding potential selection on synonymous and non-synonymous substitutions. Most human cells harbor hundreds of mtDNA molecules. Repeated encounters with a variant in a heteroplasmic state, where the variant is not found in all of the sequenced mtDNA molecules, is often considered to be a signal of pathogenicity, as homoplasmy of a deleterious variant is expected to lead to a fitness defect (Ye et al., 2014). However, if a synonymous change to mtDNA is neutral, whether a synonymous variant is encountered as heteroplasmic or homoplasmic should be a function of drift and the number of cell divisions since the initial appearance of the novel mutation in a population (Chinnery et al., 2000; Schaack et al., 2020; Stewart and Chinnery, 2015). I plotted the frequency of heteroplasmy calls against the number of samples harboring the selected variant for those substitutions encountered in at least ten HelixMTdb samples. Synonymous variants decreased in the frequency at which they were labelled as heteroplasmic as the population frequency of encounters increased, consistent either with neutral drift toward homoplasmy or with selection (Fig. 4). However, a trend toward a higher frequency of heteroplasmy calls for non-synonymous variants than for synonymous variants was easily visualized at lower population frequencies, again consistent with a general lack of selection on human synonymous variation.

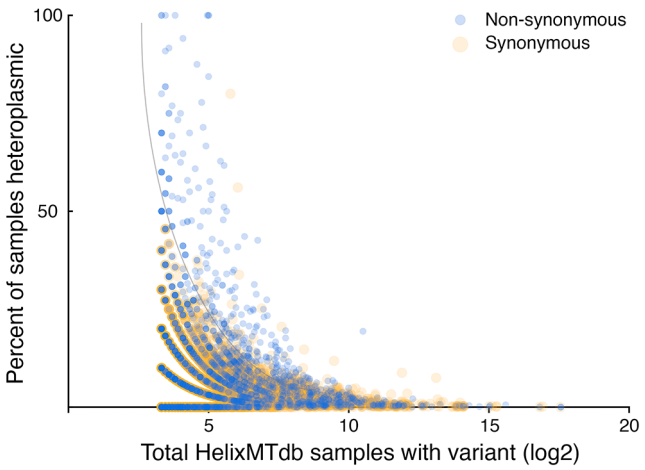

Fig. 4. Reduced selection on synonymous substitution compared to non-synonymous substitution is indicated by an analysis of variant heteroplasmy. Only variants represented by at least ten samples in HelixMTdb are plotted. The grey curve imposed upon the figure highlights a notable divergence in heteroplasmic sample fractions at low variant frequency that becomes apparent when comparing synonymous and non-synonymous variants.

DISCUSSION

Taken together, my findings indicate that human variation at mtDNA-encoded P3s is mostly constrained by the substitution rates of each nucleotide. More specifically, variation appears highly restricted by the reduced likelihood of transversion relative to transition in mtDNA, a phenomenon well documented in earlier studies of humans and other mammals (Aquadro and Greenberg, 1983; Belle et al., 2005; Brown and Simpson, 1982; Kennedy et al., 2013; Kumar, 1996; Tamura and Nei, 1993; Vermulst et al., 2007; Wakeley, 1996; Zaidi et al., 2019). Any role for selection at synonymous sites is relatively minor, with strength of selection potentially dependent upon the specific amino acid under analysis. The low selection on codon choice encountered during my study is quite consistent with earlier, more limited analyses of mitochondrial codon choice (Castellana et al., 2011; Faith and Pollock, 2003; Jia and Higgs, 2008; Uddin and Chakraborty, 2017) and with the highly streamlined tRNA set used during mitochondrial protein synthesis.

The strong link that I have revealed between mutational biases and human variant frequencies in the large HelixMTdb dataset highlights the difficulties in assigning potential pathogenicity to non-synonymous variants based, in part, upon the variant frequency in the population (Bomba et al., 2017; Gibson, 2012; McInnes et al., 2021; Povysil et al., 2019; Zuk et al., 2014). Accordingly, attempts to link rare and *de novo* variation to disease are likely to be most successful when the mutational biases for each nucleotide can be estimated and properly taken into account.

MATERIALS AND METHODS

Calculation of protein changes caused by single nucleotide substitutions from the human reference sequence

The human reference mtDNA sequence and accompanying annotation (accession NC_012920.1) was downloaded from GenBank and used as input for the script 'amino_acid_changes_caused_by_mtDNA_nucleotide_changes_in_reference.py'. This script generates a file reporting the protein sequence change associated with each simulated single base substitution to each human mitochondrial protein coding gene.

Analysis of merged human mtDNA variation

The HelixMTdb dataset (Bolze et al., 2019 preprint) was downloaded on August 26, 2021 (https://www.helix.com/pages/mitochondrial-variant-database). The script 'shape_HelixMTdb.py' takes as input the HelixMTdb dataset and the output of 'amino_acid_changes_caused_by_mtDNA_nucleotide_changes_in_reference.py', then links HelixMTdb variation to simulated amino acid changes. A subsequent analysis of transitions and transversions from the reference sequence was performed by

running the script 'transitions_transversions.py' on the output of 'shape_HelixMTdb.py'.

To check the diversity of mtDNA sequences tabulated within HelixMTdb, and to further explore this dataset, I used the script 'check_HelixMTdb_diversity.py', which tests how many silent P3 changes from the reference sequence can be identified for each amino acid position, further classified by twofold or fourfold P3 degeneracy. To simplify this analysis, leucine and serine codons were excluded from this analysis, since these amino acids are associated with tRNAs of both twofold and fourfold P3 degeneracy.

Detection of third codon position selection by analysis of mammalian mtDNAs

Records for mammalian reference mtDNAs were obtained using the Organelle Genome Resources provided by the National Center for Biotechnology Information Reference Sequence project (NCBI RefSeq, Release 207, https://www.ncbi.nlm.nih.gov/genome/organelle/) (O'Leary et al., 2016). All accessions not containing 'NC_' at the beginning of their accession name were removed, and this list of accessions was used to download full GenBank records ['mammalian_mtDNA_NC_AUG_26_2021_noroot.gb'] using the NCBI Batch Entrez server (https://www.ncbi.nlm.nih.gov/sites/batchentrez). These GenBank records were analyzed by the script 'third_codon_position_selections_all_mammals.py' to determine total counts and frequencies of P3 bases for each amino acid. Records not containing all of the following coding sequence annotations were discarded: 'ND1', 'ND2', 'COX1', 'COX2', 'ATP8', 'ATP6', 'COX3', 'ND3', 'ND4L', 'ND4', 'ND5', 'ND6', 'CYTB'.

To calculate TSSs (Akpınar et al., 2021 preprint) for I-P3 positions, the GenBank record for *I. iguana* (NC_002793.1) was added to the set of mammalian GenBank records ['mammalian_mtDNA_NC_AUG_26_2021_Iguana_root.gb']. The script 'I-P3_Part_1.py' was used to extract and align sequences from the resulting input GenBank file. Again, accessions without all of the following coding sequence annotations were discarded by this script: 'ND1', 'ND2', 'COX1', 'COX2', 'ATP8', 'ATP6', 'COX3', 'ND3', 'ND4L', 'ND4', 'ND5', 'ND6', 'CYTB'. Accessions with any coding sequence found duplicated in another accession were also discarded. This script calls upon MAFFT v7.487 (Katoh and Standley, 2013) to align mtDNA-derived coding sequences using the FFT-NS-2 algorithm. A concatenated alignment of coding sequences output from this script was used to infer a maximum likelihood tree in RAxML-NG v1.0.3 (Kozlov et al., 2019) using a single partition, a GTR+FO+G4 m model of DNA change, and a seed of 777. Ten random and ten parsimony-based starting trees were used to initiate tree construction, and the average relative Robinson-Foulds distance (Robinson and Foulds, 1981) for the inferred trees was 0.01. 600 bootstrap replicates were generated using RAxML-NG v1.0.3, and a weighted Robinson-Foulds distance converged below a 1% cutoff value (seed of 2000). Felsenstein's Bootstrap Proportions (Felsenstein, 1985) ['FBP_AUG_26_2021.raxml.support] and the Transfer Bootstrap Expectations (Lemoine et al., 2018) ['TBE_AUG_26_2021.raxml.support'] were calculated and used to label the best scoring tree ['T3_AUG_26_P3_REV.raxml.bestTree']. This tree was used for downstream analyses after using FigTree 1.4.4 (https://github.com/rambaut/figtree/releases) to place the root upon the branch leading to *I. iguana* ['AUG_26_P3_REV_tree_rooted_Iguana_iguana.nwk'].

This rooted tree and the coding sequence alignments for each mitochondria-encoded protein were used to determine the TSSs associated with each class of P3 and to determine which bases can occupy mammalian I-P3sites. Most fourfold degenerate P3s within codons with identical first and second positions were analyzed using the 'I-P3_Part_2_AGPRTV.py' script, most twofold degenerate purine P3s were analyzed with 'I-P3_Part_2_EKMQW.py', and most twofold degenerate pyrimidine P3s were analyzed with 'I-P3_Part_2_CDFHINY.py'. Any amino acids positions encoded by tRNA L1 in all mammals and the outgroup, and therefore harboring T and T at the first and second codon positions in all samples, were sought by script 'I-P3_Part_2_L1.py'. Similarly, amino acids positions encoded by tRNA L2 in all mammals and the outgroup, and therefore harboring C and T at the first and second codon positions in all samples, would have been identified and analyzed by script 'I-P3_Part_2_L2.py'. The twofold degenerate P3s associated with the use of tRNA S2 were analyzed by script 'I-P3_Part_2_S2.py', and the fourfold degenerate P3 associated with serines encoded by tRNA S1 were analyzed by script 'I-P3_Part_2_S1.py'. Within each of these scripts, MAFFT v7.487 was used to perform alignments using the FFT-NS-2 algorithm, script 'ungap_on_reference.py' v1.0 (Dunn, 2021) was used to ungap alignments based on the human reference sequence, ancestral character predictions were made at internal nodes using RAxML-NG v1.0.3 (Kozlov et al., 2019), and seqkit v0.16.1 (Shen et al., 2016) was used while formatting the node names associated with ancestral sequences.

Comparison of total substitution scores to HelixMTdb dataset

Output of the above-mentioned scripts was combined into new tables and further annotated based upon whether the P3 data were obtained from twofold degenerate purine sites ['two_fold_AG_P3_ALL.csv'], twofold degenerate pyrimidine sites ['two_fold_CT_P3_ALL.csv'], or fourfold degenerate sites ['four_fold_P3_ALL.csv']. Further processing to quantify any relationship between TSS and general P3 type, TSS and amino acid, the link between (non-)synonymous mutation and sample counts in HelixMTdb was performed using the script 'I-P3_Part_3.py'.

Statistical analyses

The Kolmogorov–Smirnov test, a non-parametric approach testing the hypothesis that two samples are drawn from different populations, was used to compare the total variant counts of synonymous and non-synonymous substitutions, to contrast total transversion versus transition counts for specific amino acids, and to compare TSS distributions between amino acids or frequency classes. Kolmogorov–Smirnov analyses were performed in Prism 9.2.0 or SciPy v1.7.1 (Virtanen et al., 2020). When analyzing I-P3 TSS values among different frequency classes, TSS distributions are arranged by amino acid, which are then compared by Kolmogorov–Smirnov testing only within a given category (fourfold degenerate, twofold degenerate purine, twofold degenerate pyrimidine). Moreover, the 'common' frequency class was not subject to statistical testing due to the limited number of variants found within this category. Correction for multiple testing (Bonferroni correction) was accomplished by multiplying each single test *P*-value by the number of tests.

Acknowledgements

I appreciate the computational support provided by the Center for Scientific Computing, Finland (Puhti supercomputer). I thank Anı Akpınar for assistance with HelixMTdb processing and for critical comments on the analysis. I also thank Gülayşe İnce Dunn and Svetlana Konovalova for helpful manuscript comments.

Competing interests

The authors declare no competing or financial interests.

Author contributions

Conceptualization: C.D.D.; Methodology: C.D.D.; Software: C.D.D.; Validation: C.D.D.; Formal analysis: C.D.D.; Investigation: C.D.D.; Data curation: C.D.D.; Writing - original draft: C.D.D.; Writing - review & editing: C.D.D.; Visualization: C.D.D.; Project administration: C.D.D.; Funding acquisition: C.D.D.

Funding

Funding for this project was obtained from the Sigrid Jusélius Foundation (Senior Researcher Grant to C.D.D.) and the European Research Council (ERC Starting Grant RevMito 637649 to C.D.D.).

Data availability

The software and data that support the findings of this study are available at: https://doi.org/10.5281/zenodo.5493479.

References

Akpınar, B. A., Sharma, V. and Dunn, C. D. (2021). A novel approach to the detection of unusual mitochondrial protein change suggests hypometabolism of ancestral simians. *bioRxiv*. doi:10.1101/2021.03.10.434614

Aquadro, C. F. and Greenberg, B. D. (1983). Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. *Genetics* **103**, 287-312. doi:10.1093/genetics/103.2.287

Belle, E. MS., Piganeau, G., Gardner, M. and Eyre-Walker, A. (2005). An investigation of the variation in the transition bias among various animal mitochondrial DNA. *Gene* **355**, 58-66. doi:10.1016/j.gene.2005.05.019

Bolze, A., Mendez, F., White, S., Tanudjaja, F., Isaksson, M., Jiang, R., Dei Rossi, A., Cirulli, E. T., Rashkin, M., Metcalf, W. J. et al. (2019). A catalog of homoplasmic and heteroplasmic mitochondrial DNA variants in humans. *bioRxiv*. doi:10.1101/798264

Bomba, L., Walter, K. and Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* **18**, 77. doi:10.1186/s13059-017-1212-4

Brown, G. G. and Simpson, M. V. (1982). Novel features of animal mtDNA evolution as shown by sequences of two rat cytochrome oxidase subunit II genes. *Proc. Natl. Acad. Sci. USA* **79**, 3246-3250. doi:10.1073/pnas.79.10.3246

Castellana, S., Vicario, S. and Saccone, C. (2011). Evolutionary patterns of the mitochondrial genome in Metazoa: exploring the role of mutation and selection in mitochondrial protein coding genes. *Genome Biol Evol* **3**, 1067-1079. doi:10.1093/gbe/evr040

Chinnery, P. F., Thorburn, D. R., Samuels, D. C., White, S. L., Dahl, H. M., Turnbull, D. M., Lightowlers, R. N. and Howell, N. (2000). The inheritance of mitochondrial DNA heteroplasmy: random drift, selection or both? *Trends Genet.* **16**, 500-505. doi:10.1016/S0168-9525(00)02120-X

Claussnitzer, M., Cho, J. H., Collins, R., Cox, N. J., Dermitzakis, E. T., Hurles, M. E., Kathiresan, S., Kenny, E. E., Lindgren, C. M., MacArthur, D. G. et al. (2020). A brief history of human disease genetics. *Nature* **577**, 179-189. doi:10.1038/s41586-019-1879-7

Dunn, C. D. (2021). Ungap_on_reference_v_1_0. doi:10.5281/zenodo.4633159

Faith, J. J. and Pollock, D. D. (2003). Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics* **165**, 735-745. doi:10.1093/genetics/165.2.735

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783-791. doi:10.1111/j.1558-5646.1985.tb00420.x

Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D. J. et al. (2016). The genetic architecture of type 2 diabetes. *Nature* **536**, 41-47. doi:10.1038/nature18642

Genovese, G., Fromer, M., Stahl, E. A., Ruderfer, D. M., Chambert, K., Landén, M., Moran, J. L., Purcell, S. M., Sklar, P., Sullivan, P. F. et al. (2016). Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat. Neurosci.* **19**, 1433-1441. doi:10.1038/nn.4402

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135-145. doi:10.1038/nrg3118

Gorman, G. S., Chinnery, P. F., DiMauro, S., Hirano, M., Koga, Y., McFarland, R., Suomalainen, A., Thorburn, D. R., Zeviani, M. and Turnbull, D. M. (2016). Mitochondrial diseases. *Nat. Rev. Dis. Primers* **2**, 16080. doi:10.1038/nrdp.2016.80

Ingman, M., Kaessmann, H., Pääbo, S. and Gyllensten, U. (2000). Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708-713. doi:10.1038/35047064

Jia, W. and Higgs, P. G. (2008). Codon usage in mitochondrial genomes: distinguishing context-dependent mutation from translational selection. *Mol. Biol. Evol.* **25**, 339-351. doi:10.1093/molbev/msm259

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780. doi:10.1093/molbev/mst010

Kennedy, S. R., Salk, J. J., Schmitt, M. W. and Loeb, L. A. (2013). Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. *PLoS Genet.* **9**, e1003794. doi:10.1371/journal.pgen.1003794

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. and Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453-4455. doi:10.1093/bioinformatics/btz305

Kumar, S. (1996). Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates. *Genetics* **143**, 537-548. doi:10.1093/genetics/143.1.537

Lemoine, F., Domelevo Entfellner, J.-B., Wilkinson, E., Correia, D., Dávila Felipe, M., De Oliveira, T. and Gascuel, O. (2018). Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556**, 452-456. doi:10.1038/s41586-018-0043-0

Lencz, T., Yu, J., Khan, R., Lam, M., Flaherty, E., Maniatis, T., Malhotra, A., Atzmon, G. and Pe'er, I. (2021). Ultra-rare exonic variants identified in a founder population implicate cadherins and protocadherins in schizophrenia. *Biol. Psychiatry* **109**, 1465-1478.e4. doi:10.1016/j.biopsych.2021.02.222

Luo, Y., de Lange, K. M., Jostins, L., Moutsianas, L., Randall, J., Kennedy, N. A., Lamb, C. A., McCarthy, S., Ahmad, T., Edwards, C. et al. (2017). Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. *Nat. Genet.* **49**, 186-192. doi:10.1038/ng.3761

Maca-Meyer, N., González, A. M., Larruga, J. M., Flores, C. and Cabrera, V. M. (2001). Major genomic mitochondrial lineages delineate early human expansions. *BMC Genet.* **2**, 13. doi:10.1186/1471-2156-2-13

Macklin, S., Durand, N., Atwal, P. and Hines, S. (2018). Observed frequency and challenges of variant reclassification in a hereditary cancer clinic. *Genet. Med.* **20**, 346-350. doi:10.1038/gim.2017.207

Manrai, A. K., Funke, B. H., Rehm, H. L., Olesen, M. S., Maron, B. A., Szolovits, P., Margulies, D. M., Loscalzo, J. and Kohane, I. S. (2016). Genetic Misdiagnoses and the Potential for Health Disparities. *N. Engl. J. Med.* **375**, 655-665. doi:10.1056/NEJMsa1507092

McInnes, G., Sharo, A. G., Koleske, M. L., Brown, J. E. H., Norstad, M., Adhikari, A. N., Wang, S., Brenner, S. E., Halpern, J., Koenig, B. A. et al. (2021). Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. *Am. J. Hum. Genet.* **108**, 535-548. doi:10.1016/j.ajhg.2021.03.003

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D. et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-D745. doi:10.1093/nar/gkv1189

Povysil, G., Petrovski, S., Hostyk, J., Aggarwal, V., Allen, A. S. and Goldstein, D. B. (2019). Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat. Rev. Genet.* **20**, 747-759. doi:10.1038/s41576-019-0177-4

Reyes, A., Gissi, C. and Pesole, G. (1998). Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol Biol* **15**, 957-966. doi:10.1093/oxfordjournals.molbev.a026011

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131-147. doi:10.1016/0025-5564(81)90043-2

Sazonovs, A. and Barrett, J. C. (2018). Rare-variant studies to complement genome-wide association studies. *Annu. Rev. Genomics Hum. Genet.* **19**, 97-112. doi:10.1146/annurev-genom-083117-021641

Schaack, S., Ho, E. KH. and Macrae, F. (2020). Disentangling the intertwined roles of mutation, selection and drift in the mitochondrial genome. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190173. doi:10.1098/rstb.2019.0173

Shendure, J., Findlay, G. M. and Snyder, M. W. (2019). Genomic Medicine-Progress, Pitfalls, and Promise. *Cell* **177**, 45-57. doi:10.1016/j.cell.2019.02.003

Shen, W., Le, S., Li, Y. and Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* **11**, e0163962. doi:10.1371/journal.pone.0163962

Stewart, J. B. and Chinnery, P. F. (2015). The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nat. Rev. Genet.* **16**, 530-542. doi:10.1038/nrg3966

Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512-526. doi:10.1093/oxfordjournals.molbev.a040023

Thompson, K., Collier, J. J., Glasgow, R. IC., Robertson, F. M., Pyle, A., Blakely, E. L., Alston, C. L., Oláhová, M., McFarland, R. and Taylor, R. W. (2020). Recent advances in understanding the molecular genetic basis of mitochondrial disease. *J. Inherit. Metab. Dis.* **43**, 36-50. doi:10.1002/jimd.12104

Uddin, A. and Chakraborty, S. (2017). Synonymous codon usage pattern in mitochondrial CYB gene in pisces, aves, and mammals. *Mitochondrial DNA A DNA Mapp Seq Anal* **28**, 187-196. doi:10.3109/19401736.2015.1115842

Uricchio, L. H., Zaitlen, N. A., Ye, C. J., Witte, J. S. and Hernandez, R. D. (2016). Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants. *Genome Res.* **26**, 863-873. doi:10.1101/gr.202440.115

Vermulst, M., Bielas, J. H., Kujoth, G. C., Ladiges, W. C., Rabinovitch, P. S., Prolla, T. A. and Loeb, L. A. (2007). Mitochondrial point mutations do not limit the natural lifespan of mice. *Nat. Genet.* **39**, 540-543. doi:10.1038/ng1988

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J. et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261-272. doi:10.1038/s41592-019-0686-2

Wakeley, J. (1996). The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends Ecol. Evol.* **11**, 158-162. doi:10.1016/0169-5347(96)10009-4

Walker, J. E., Carroll, J., Altman, M. C. and Fearnley, I. M. (2009). *Chapter 6 Mass Spectrometric Characterization of the Thirteen Subunits of Bovine Respiratory Complexes that are Encoded in Mitochondrial DNAMethods in Enzymology*, pp. 111-131. Academic Press. doi:10.1016/S0076-6879(08)04406-6

Ye, K., Lu, J., Ma, F., Keinan, A. and Gu, Z. (2014). Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals. *Proc. Natl. Acad. Sci. USA* **111**, 10654-10659. doi:10.1073/pnas.1403521111

Zaidi, A. A., Wilton, P. R., Su, M. S.-W., Paul, I. M., Arbeithuber, B., Anthony, K., Nekrutenko, A., Nielsen, R. and Makova, K. D. (2019). Bottleneck and selection in the germline and maternal age influence transmission of mitochondrial DNA in human pedigrees. *Proc. Natl. Acad. Sci. USA* **116**, 25172-25178. doi:10.1073/pnas.1906331116

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R. and Lander, E. S. (2014). Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. USA* **111**, E455-E464. doi:10.1073/pnas.1322563111