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

Cory D. Dunn

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

Contact details:

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

Abstract:

Genome-wide association studies (GWASs) typically seek common genetic variants that can influence disease likelihood. However, these analyses often fail to convincingly link specific genes and their variants with highly penetrant phenotypic effects [1–4]. To solve the 'missing heritability problem' that characterizes GWASs, researchers have turned to rare variants revealed by next-generation sequencing when seeking genomic changes that may be pathogenic, as a reduction in variant frequency is an expected outcome of selection. While triage of rare variants has led to some success in illuminating genes linked to heritable disease [5–8], the interpretation and utilization of rare genomic changes remains very challenging [9,10].

Human mitochondrial DNA (mtDNA) encodes proteins and RNAs required for the essential process of oxidative phosphorylation, and a number of metabolic diseases are linked to mitochondrial mutations [11,12]. 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 [13]. Here, we were surprised to find that many synonymous nucleotide substitutions were never detected within this quite substantial survey of human mtDNA. Subsequent study of more than 1000 mammalian mtDNAs suggested that selection on synonymous sites within mitochondrial protein-coding genes is minimal and unlikely to explain the rarity of most synonymous changes among humans. Rather, the mutational propensities of mtDNA are more likely to determine variant frequency. Our findings have general implications for the interpretation of variant frequencies when studying heritable disease.

Main:

During an exploration of selective pressures that may act upon mitochondria-encoded polypeptides, we simulated every single nucleotide substitution from the human reference mtDNA sequence within all protein-coding sequences, then cross-referenced these simulated changes with those found in HelixMTdb. Consistent with selection acting upon most amino acid changes, non-synonymous substitutions were depleted to a greater extent than synonymous substitutions when considering either the number of samples harboring each variant (Fig. 1a) or whether the nucleotide substitution was encountered at all during generation of the HelixMTdb (Fig. 1b).

Since synonymous changes would not be expected to change the structure or function of mitochondria-encoded proteins, we were surprised by the sheer 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. We tested whether SSNEs were predominantly associated with specific amino acids, and we found that the SSNE fraction of synonymous changes was substantially larger for the amino acids arginine, threonine, alanine, valine, serine, glycine, proline, and leucine than for the other amino acids (Fig. 1c). All of these amino acids can be encoded by codons exhibiting four-fold degeneracy, or the ability to accommodate any nucleotide change at the third codon position (here, 'P3') without changing the translation product. Other amino acids, in contrast, are restricted to only two P3 nucleotides, with synonymous change occurring only by transition (a purine-purine change or a pyrimidine-pyrimidine change) and not by transversion (a purine-pyrimidine change, or *vice versa*).

We tested whether SSNEs might be more closely associated with transversions or transitions when considering these amino acids for which P3 can be four-fold degenerate, and we found that nearly all SSNEs assigned to these eight amino acids (> 96%) were linked to a potential transversion (Fig. 1d). Expanding our 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 four-fold degenerate P3s was a transition or a transversion (Fig. 1e).

Two conspicuous and non-exclusive hypotheses exist regarding the notable enrichment of transversions among SSNEs at four-fold degenerate P3s. First, there may be unanticipated and strong 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. To address the first possibility, we 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 1244 mammalian mtDNAs. Here, we have also taken into account the two different mitochondrial tRNAs recognizing leucine codons and the two mitochondrial tRNAs recognizing serine codons. As encountered in previous studies [14,15] , guanine was depleted from mitochondrial P3s for which the presence of any purine does not lead to an amino acid change (Fig. 2a), 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 when considering the relative depletion of guanine from P3s, guanine was nonetheless detected at more than 3000 P3 positions accepting synonymous purine substitution when considering every amino acid (Fig. 2b). Therefore, nucleotide frequencies at P3s are unlikely to reflect significant, widespread selection on codons that would be inherent to the process of mitochondrial translation, a result consistent with earlier, more limited analyses of mitochondrial codon choice [16–19] and with the highly streamlined tRNA set available for mitochondrial protein synthesis.

While degeneracy at the third codon position is a general feature of mtDNA-encoded amino acids, these results are not necessarily informative about the possibility that substantial selection acts upon *specific* P3s encoded by the mitochondrial genome. To further explore the extent to which individual P3s might be under selection, we focused our attention upon codons for which the first and second positions, and therefore the encoded amino acids, are 100% identical in an alignment consisting of 1179 mammals and an outlier reptile sequence (*Anolis punctatus*) 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 protein alignment sites harboring only leucine. We found that 564/565 (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. Only the I-P3 of the codon annotated as the COX3 starting methionine in humans appears to be totally constrained with respect to nucleotide choice, as this P3 is always occupied by guanine in mammals. Interestingly, COX3 is the only mitochondrial polypeptide which lacks a formyl-methionine at its amino-terminus [20], although whether there is a mechanistic relationship between these two observations remains to be determined.

Next, we calculated the Total Substitution Score (TSS [21]), or the number of substitutions at a given site occurring throughout our inferred mammalian phylogenetic tree, for each I-P3. We found, with few exceptions, that nearly all four-fold degenerate, two-fold degenerate pyrimidine, and two-fold degenerate purine I-P3s have been subject to synonymous substitution tens, or even hundreds of times, during approximately 200 million years of mammal evolution (Fig. 2c), a result quite consistent with minimal selection upon nucleotide choice at mitochondrial P3s. However, we did note statistically significant divergence between TSS distributions at I-P3s when comparing amino acids within a given degeneracy class, indicating that selection may act upon at least some mitochondrial third codon positions.

Next, we asked whether a low frequency of human variants at four-fold degenerate I-P3s would correspond with lower TSS values of corresponding P3 positions, a potential indicator of selection extending to humans. Here, we 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, we detected no significant relationship between TSS and variant frequency for four-fold degenerate I-P3s (Fig. 2d), indicating that the highly elevated SSNE abundance at four-fold degenerate P3s is unlikely to be due to selection. For two-fold degenerate purine and two-fold degenerate pyrimidine I-P3s, we found evidence of mammal-wide selection that might determine the frequency of some variants in humans. 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 and when comparing TSS distributions between absent and rare variants at I-P3s accepting pyrimidines.

Finally, we explored how the quantification of heteroplasmic samples provided by HelixMTdb might be informative regarding potential selection on synonymous substitutions. Hundreds of mtDNA molecules are found within most human cells. 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 [22]. However, if a synonymous change to mtDNA is neutral, whether a synonymous variant is encountered as heteroplasmic or homoplasmic should be the result of drift and a function of the number of cell divisions since the initial appearance of the novel mutation in a population [23–25]. We plotted the frequency of heteroplasmy calls against the number of samples harboring the selected variant for those variants encountered in at least 10 HelixMTdb samples. Synonymous variants decreased in the frequency at which they were detected as heteroplasmic as the population frequency of encounters increased, consistent either with neutral drift toward homoplasmy or with selection (Fig. 2e). 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 synonymous variation.

Conclusions and outlook:

Taken together, our 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 result supported by earlier studies of humans and other mammals [15,26–35]. Any role for selection at synonymous sites is relatively minor, with strength of selection potentially dependent upon the specific amino acid under analysis. The strong link that we 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 [1,3,4,36,37] . 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.

Data availability:

The software and data that support the findings of this study are available at https://github.com/corydunnlab/human_mito_variation.

Methodology:

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 ['NC_012920.gbk'] and used as input for the script 'amino_acid_changes_caused_by_mtDNA_nucleotide_changes_in_reference.py'.

Analysis of merged human mtDNA variation

The HelixMTdb dataset [13] was downloaded on April 8, 2021. The script 'shape_HelixMTdb.py' was run using the output of 'amino_acid_changes_caused_by_mtDNA_nucleotide_changes_in_reference.py' and the HelixMTdb dataset. An analysis of transitions and transversions from the reference sequence was performed by running the script 'transitions_transversions.py' using the output of 'shape_HelixMTdb.py'.

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 204, https://www.ncbi.nlm.nih.gov/genome/organelle/) [38]. All accessions not containing 'NC_' at the beginning of their accession name were removed, and this list accessions was used to download full GenBank records using the NCBI Batch Entrez server (https://www.ncbi.nlm.nih.gov/sites/batchentrez). ['mito_synonymous_mammals_WO_A_punctatus.gb']. 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 [21] for I-P3 positions, the GenBank record for *Anolis punctatus* (NC_044125.1) was added to the set of mammalian GenBank records ['mito_synonymous_mammals_AND_A_punctatus.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 v7.475 [39] 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 v1.0.2 [40] using a single partition, a GTR+FO+G4m model of DNA change, and a seed of 566. 10 random and 10 parsimony-based starting trees were used to initiate tree construction, and the average relative Robinson-Foulds distance [41] for the inferred trees was 0.01. 600 bootstrap replicates were generated using RAxML-NG v1.0.2, and a weighted Robinson-Foulds distance converged below a 1% cutoff value. Felsenstein's Bootstrap Proportions [42] ['FBP_mito_synonymous.raxml.support] and the Transfer Bootstrap Expectations [43] ['TBE_mito_synonymous.raxml.support'] were calculated and used to label our best scoring tree ['mito_synonymous.raxml.bestTree']. The best maximum likelihood 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 *Anolis punctatus* ['mito_synonymous_best_tree_rooted_Anolis_punctatus.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. Most four-fold degenerate P3s within codons with identical first and second positions were analyzed using the 'I-P3_Part_2_AGPRTV.py' script, most two-fold degenerate purine P3s were analyzed with 'I-P3_Part_2_EKMQW.py', and most two-fold degenerate pyrimidine P3s were analyzed with 'I-P3_Part_2_CDFHINY.py'. However, 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, were would have been identified and analyzed by script 'I-P3_Part_2_L2.py'. The two-fold degenerate P3s associated with the use of tRNA S2 were analyzed by script 'I-P3_Part_2_S2.py', and the four-fold 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 v7.475 was used for alignments using the FFT-NS-2 algorithm, script 'ungap_on_reference.py' v1.0 [44] was used to ungap alignments based on the human reference sequence, ancestral character predictions were made at internal nodes using RAxML-NG v1.0.2 [40], and seqkit v0.160 [45] was used for manipulating 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 two-fold degenerate purine sites ['two_fold_AG_P3_TSS.csv'], two-fold degenerate pyrimidine sites ['two_fold_CT_P3_TSS.csv'], or four-fold degenerate sites ['four_fold_P3_TSS.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

Kolmogorov-Smirnov analysis of the total variant counts of synonymous and non-synonymous substitutions and Kolmogorov-Smirnov analysis of total transversion versus transition counts for specific amino acids were performed in Prism 9.1.0. Kolmogorov-Smirnov comparisons of amino acid or variant frequency class with TSS were carried out using SciPy v1.6.0 [46]. Here, the 'common' frequency class was not subject to statistical testing due to a maximum of only two common variants within each I-P3 class. Correction for multiple testing (Bonferroni correction) was accomplished by multiplying each single test P-value by the number of tests.

Acknowledgements:

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.). We appreciate the essential computational support provided by the Center for Scientific Computing, Finland (Puhti supercomputer), as well as assistance from Anı Akpınar with processing the HelixMTdb. We thank Gülayşe İnce Dunn and Svetlana Konovalova for helpful comments on the manuscript.

References:

1. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* **18**, 77 (2017).
2. Sazonovs, A. & Barrett, J. C. Rare-Variant Studies to Complement Genome-Wide Association Studies. *Annu. Rev. Genomics Hum. Genet.* **19**, 97–112 (2018).
3. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E455–64 (2014).
4. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135–145 (2012).
5. Lencz, T. *et al.* Ultra-Rare Exonic Variants Identified in a Founder Population Implicate Cadherins and Protocadherins in Schizophrenia. *Biological Psychiatry* vol. 89 S83 (2021).
6. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
7. Genovese, G. *et al.* Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat. Neurosci.* **19**, 1433–1441 (2016).
8. Luo, Y. *et al.* Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. *Nat. Genet.* **49**, 186–192 (2017).
9. Manrai, A. K. *et al.* Genetic Misdiagnoses and the Potential for Health Disparities. *N. Engl. J. Med.* **375**, 655–665 (2016).
10. Macklin, S., Durand, N., Atwal, P. & Hines, S. Observed frequency and challenges of variant reclassification in a hereditary cancer clinic. *Genet. Med.* **20**, 346–350 (2018).
11. Thompson, K. *et al.* Recent advances in understanding the molecular genetic basis of mitochondrial disease. *J. Inherit. Metab. Dis.* **43**, 36–50 (2020).
12. Gorman, G. S. *et al.* Mitochondrial diseases. *Nat Rev Dis Primers* **2**, 16080 (2016).
13. Bolze, A. *et al.* Selective constraints and pathogenicity of mitochondrial DNA variants inferred from a novel database of 196,554 unrelated individuals. *bioRxiv* 1151 (2019).
14. Reyes, A., Gissi, C. & Pesole, G. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol. Biol.* (1998).
15. Kumar, S. Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates. *Genetics* **143**, 537–548 (1996).
16. Jia, W. & Higgs, P. G. Codon usage in mitochondrial genomes: distinguishing context-dependent mutation from translational selection. *Mol. Biol. Evol.* **25**, 339–351 (2008).
17. Castellana, S., Vicario, S. & Saccone, C. 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 (2011).
18. Uddin, A. & Chakraborty, S. Synonymous codon usage pattern in mitochondrial CYB gene in pisces, aves, and mammals. *Mitochondrial DNA A DNA Mapp Seq Anal* **28**, 187–196 (2017).
19. Faith, J. J. & Pollock, D. D. Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics* **165**, 735–745 (2003).
20. Walker, J. E., Carroll, J., Altman, M. C. & Fearnley, I. M. Chapter 6 Mass Spectrometric Characterization of the Thirteen Subunits of Bovine Respiratory Complexes that are Encoded in Mitochondrial DNA. in *Methods in Enzymology* vol. 456 111–131 (Academic Press, 2009).
21. Akpınar, B. A., Sharma, V. & Dunn, C. D. A novel approach to the detection of unusual mitochondrial protein change suggests hypometabolism of ancestral simians. *bioRxiv* 2021.03.10.434614 (2021) doi:10.1101/2021.03.10.434614.
22. Ye, K., Lu, J., Ma, F., Keinan, A. & Gu, Z. Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 10654–10659 (2014).
23. Chinnery, P. F. *et al.* The inheritance of mitochondrial DNA heteroplasmy: random drift, selection or both? *Trends Genet.* **16**, 500–505 (2000).

24. Stewart, J. B. & Chinnery, P. F. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nat. Rev. Genet.* **16**, 530–542 (2015).
25. Schaack, S., Ho, E. K. H. & Macrae, F. Disentangling the intertwined roles of mutation, selection and drift in the mitochondrial genome. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190173 (2020).
26. Belle, E. M. S., Piganeau, G., Gardner, M. & Eyre-Walker, A. An investigation of the variation in the transition bias among various animal mitochondrial DNA. *Gene* **355**, 58–66 (2005).
27. Brown, G. G. & Simpson, M. V. Novel features of animal mtDNA evolution as shown by sequences of two rat cytochrome oxidase subunit II genes. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 3246–3250 (1982).
28. Wakeley, J. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends Ecol. Evol.* **11**, 158–162 (1996).
29. Vermulst, M. *et al.* Mitochondrial point mutations do not limit the natural lifespan of mice. *Nat. Genet.* **39**, 540–543 (2007).
30. Kennedy, S. R., Salk, J. J., Schmitt, M. W. & Loeb, L. A. Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. *PLoS Genet.* **9**, e1003794 (2013).
31. Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526 (1993).
32. Ju, Y. S. *et al.* Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *Elife* **3**, (2014).
33. Li, D. *et al.* Site-specific selection reveals selective constraints and functionality of tumor somatic mtDNA mutations. *J. Exp. Clin. Cancer Res.* **36**, 168 (2017).
34. Skonieczna, K. *et al.* Mitogenomic differences between the normal and tumor cells of colorectal cancer patients. *Hum. Mutat.* **39**, 691–701 (2018).
35. Zaidi, A. A. *et al.* Bottleneck and selection in the germline and maternal age influence transmission of mitochondrial DNA in human pedigrees. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 25172–25178 (2019).
36. McInnes, G. *et al.* Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. *Am. J. Hum. Genet.* **108**, 535–548 (2021).
37. Povysil, G. *et al.* Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat. Rev. Genet.* **20**, 747–759 (2019).
38. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45 (2016).
39. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
40. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
41. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
42. Felsenstein, J. CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP. *Evolution* **39**, 783–791 (1985).
43. Lemoine, F. *et al.* Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556**, 452–456 (2018).
44. Dunn, C. D. *Ungap_on_reference_v_1_0*. (2021). doi:10.5281/zenodo.4633159.
45. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* **11**, e0163962 (2016).
46. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

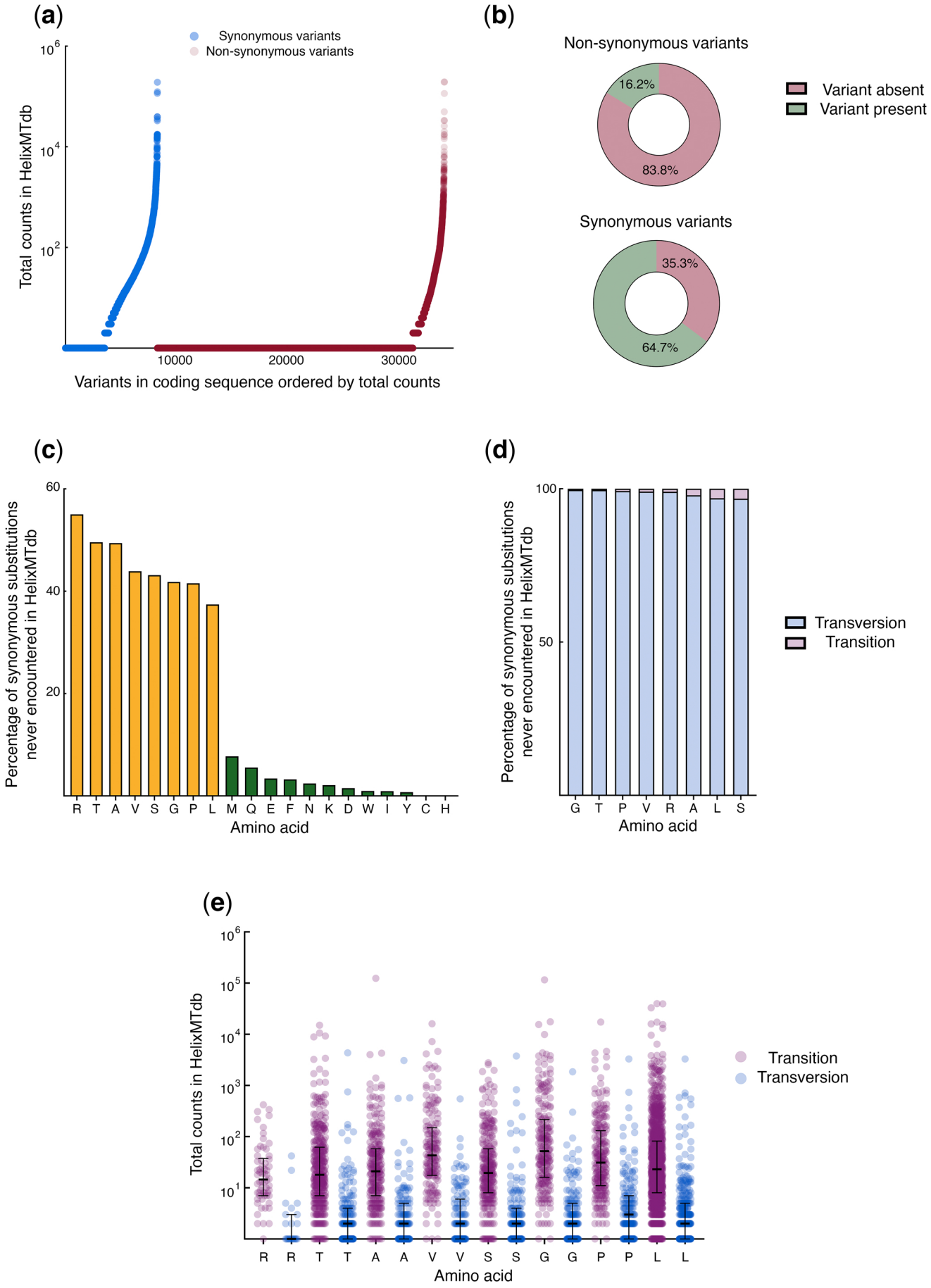

Fig. 1: Transversion substitutions are heavily depleted at four-fold degenerate third-codon positions of mitochondrial coding sequences. (**a**) Non-synonymous and synonymous variants differ in their population frequency. The sample count for variants within coding regions that are reachable by a single substitution from the human mtDNA reference sequence were obtained from HelixMTdb and plotted (Kolmogorov-Smirnov approximate P-value, <0.0001). (**b**) While the great majority of non-synonymous substitutions were never encountered by the HelixMTdb study, a substantial fraction of synonymous substitutions are also apparently lacking from the human population. (**c**) Synonymous mutations never encountered in the HelixMTdb study are most abundant at amino acids for which at least one codon is four-fold degenerate at P3 (amino acids with four-fold degenerate P3s, orange; other amino acids, green). (**d**) The vast majority of synonymous substitutions never encountered by the HelixMTdb study are transversions. (**e**) For those substitutions 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 four-fold degenerate at P3, the population count is plotted. Bar and error bars represent median and interquartile range.

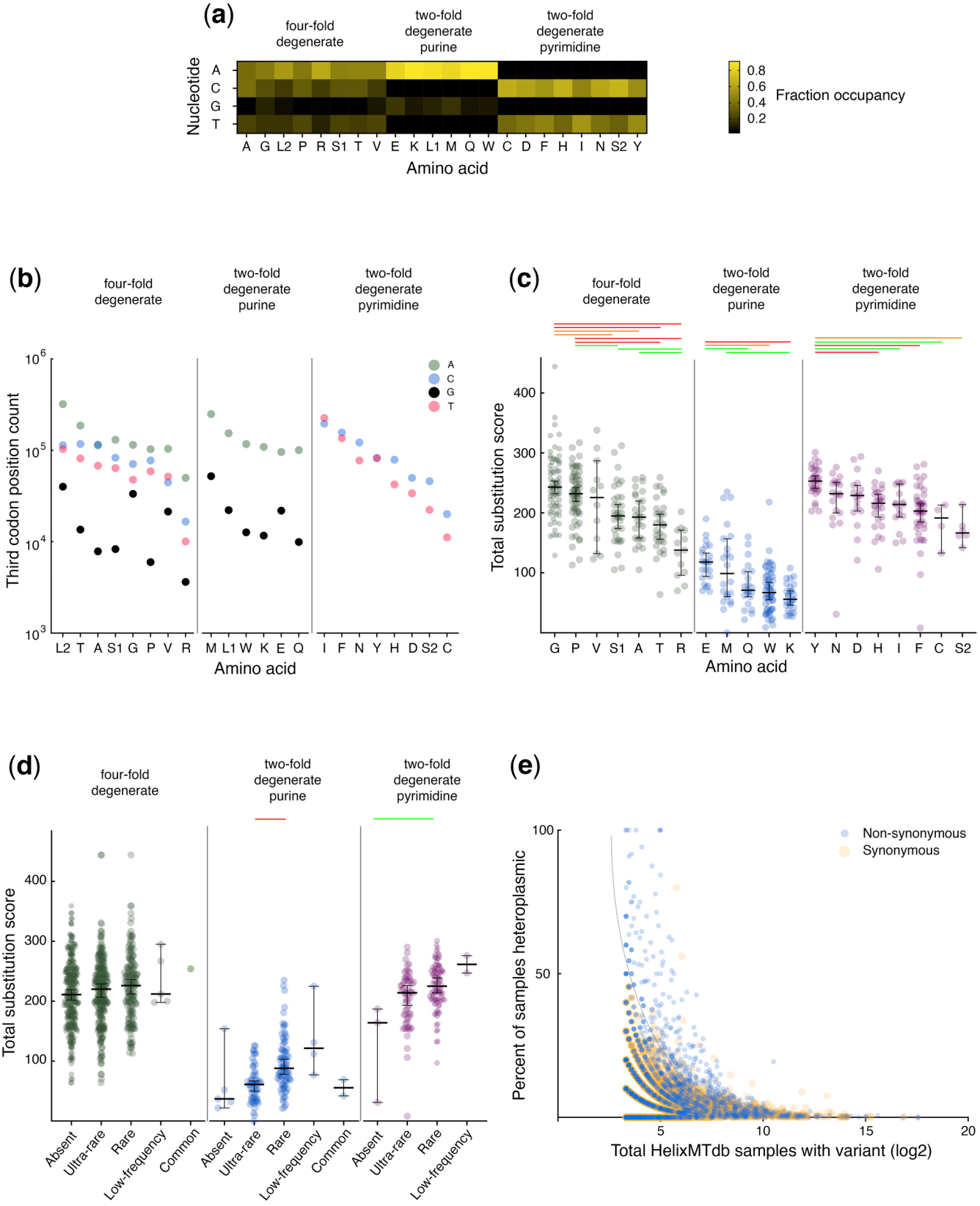

Fig. 2: Abundant substitution and degeneracy across the third codon positions of mammals. (**a**) Base occupancy is not equally distributed among nucleotides at degenerate P3s (n ≥ 31195 instances of each amino acid are analyzed across the set of input mammalian mtDNAs). (**b**) Guanine is not excluded from the P3 of any amino acid. (**c**) 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 (four-fold degenerate, two-fold degenerate purine, or two-fold 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. (**d**) Population frequency of synonymous variants is unlinked to TSS at four-fold degenerate I-P3s. Statistical significance is demonstrated as in (**c**), and the bar and error bars represent median and 95% confidence interval. (**e**) Reduced selection on synonymous substitution compared to non-synonymous substitution is indicated by an analysis of variant heteroplasmy. Only variants represented by at least 10 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.