
Supplementary information

**Quantifying constraint in the human
mitochondrial genome**

In the format provided by the
authors and unedited

Supplementary Information

Quantifying constraint in the human mitochondrial genome

Nicole J. Lake, Kaiyue Ma, Wei Liu, Stephanie L. Battle, Kristen M. Laricchia, Grace Tiao, Daniela Puiu, Kenneth K. Ng, Justin Cohen, Alison G. Compton, Shannon Cowie, John Christodoulou, David R. Thorburn, Hongyu Zhao, Dan E. Arking, Shamil R. Sunyaev & Monkol Lek

Table of Contents

Supplementary Methods	3
1. Rationale for why nuclear constraint methods could not be applied	3
2. Mitochondrial composite likelihood model	6
Supplementary Table 1: Mutation classification system used	7
3. Mitochondrial de novo mutation dataset.....	13
Supplementary Table 2: Overview of the de novo mutation datasets used	13
Supplementary Figure 1: Mutation likelihoods of de novo transitions by source	14
4. Validation of mutation model	16
Supplementary Figure 2: Assessing the predictive value of the mutational models	17
5. Assessment of mitochondrial constraint	18
Supplementary Figure 3: Confidence interval characteristics.	20
6. Simulation of germline mtDNA mutation and heteroplasmy	21
Supplementary Figure 4: Simulation of mutation and heteroplasmy in maternal lineages..	22
7. Regional constraint	24
Supplementary Figure 5: Distribution of expected values for <i>k</i> -mer by length.....	25
Supplementary Figure 6: Overview of the regional missense constraint method	26
8. Local constraint score	27
Supplementary Figure 7: Overview of the mtDNA local constraint (MLC) score.....	28
9. Comparison of constraint with traditional methods.....	29
Supplementary Figure 8: Regions with same population frequency have different mutability...30	
Supplementary Figure 9: MLC scores vary across heteroplasmies with the same frequency...31	
10. Population databases	33
Supplementary Table 3: Overview of the mitochondrial population databases analyzed	33

11. Mitochondrial base editing	34
Supplementary Table 4: mtDNA sequences recognized by TALE proteins	34
Supplementary Table 5: TALE amino acid sequences	34
Supplementary Figure 10: Bystander edits and off-target rates.....	36
Supplementary Table 6: Next generation sequencing validation of base editing	37
Supplementary Table 7: Primers used for sequencing.....	38
Supplementary Discussion.....	39
1. Extended discussion on simulation of germline mtDNA mutation and heteroplasmy	39
2. Extended discussion on missense constraint in the mitochondrial vs nuclear genome	39
3. Extended discussion on regional missense constraint within protein genes.....	40
4. Extended discussion of regional constraint within rRNA genes	41
5. Extended discussion of local constraint score across the non-coding control region.....	41
6. Extended discussion on the local constraint scores of disease-associated variation.....	42
7. Extended discussion of the association between the MSS and blood cell counts	43
8. Extended discussion on constraint in genes and regions overlooked in disease analyses	43
9. Extended discussion on selection against variants in the mtDNA.....	44
Supplementary References.....	45
Supplementary Datasets.....	50
Supplementary Video.....	52

Supplementary Methods

1. Rationale for why nuclear constraint methods could not be applied

The mitochondrial genome (mtDNA) has many unique features that differentiate it from the nuclear genome. This includes, but is not limited to, maternal inheritance, smaller size, absence of introns, high copy number, presence of heteroplasmy, germline bottleneck, distinct mutational mechanisms, and a higher rate of mutation^{1,2}. These unique features mean that the nuclear genome constraint model cannot be applied to the mtDNA, necessitating the development of an approach that is tailor-made to handle the unique features of this genome. In this section, we focus on five key reasons for why nuclear constraint methods could not be applied, providing rationale for why methods development and technical advances were required to assess mitochondrial constraint.

1.1 Why the nuclear mutational model could not be applied

Constraint is determined by comparing the observed level of genetic variation in a population to that expected under neutrality, as calculated by a mutational model. Therefore, the first step in assessing constraint is to establish a mutational model to use. Nuclear mutational models could not be applied since the two genomes have distinct mutational mechanisms and signatures. For example, the transition:transversion ratio is approximately 10-15x higher in the mitochondrial vs nuclear genome^{3,4}, and a primary source of mutations in the mitochondria are errors during replication caused by the mitochondrial DNA polymerase^{5,6}; collectively reflecting that different DNA replication and repair machinery are used in the mitochondria and nucleus⁷. Secondly, mutational models used to compute nuclear gene constraint metrics in gnomAD assign trinucleotide mutations the same rate irrespective of the strand they occur on (reference vs reverse complement)⁸. This is not a suitable assumption for the mtDNA, which has replicative strand bias for transitions⁵. Therefore, we needed to build a mitochondrial mutational model (see section 2).

1.2 Why the nuclear method for building a mutational model could not be applied

For nuclear gene constraint methods applied to gnomAD, mutation rates for each trinucleotide are computed using the non-coding regions of the genome^{8,9}. In brief, a mutation rate table is constructed using ultra-rare non-coding variants in intergenic or intronic regions (and satisfying several other criteria), from which nuclear trinucleotide mutation rates are calculated. There are

several reasons why this approach is not suitable or feasible in the mtDNA. Firstly, unlike the nuclear genome, only a small proportion (~10%) of the mtDNA is non-coding. Due to the small size of the mtDNA, this leads to insufficient counts for the 192 trinucleotide mutations across both strands (reference vs reverse complement), such that many would be assigned a non-usable mutation rate of zero. Secondly, a segment of non-coding mtDNA called the OriB-OriH region is known to have an inverted mutational signature for transitions⁵, further restricting the non-coding space available for mutational modeling. Therefore, a different approach was necessary in the mtDNA; one which also needed to handle possible sparsity of counts (see section 2).

1.3. Why the nuclear method of assessing observed:expected was not suited for mtDNA

Heteroplasmy is a phenomenon associated with the high copy number of the mtDNA, whereby a variant has a mutant load (i.e. heteroplasmy level) between 0 and 100%. Nuclear constraint models assess the observed vs expected number of unique variants in a gene or region, such that all observed variants in a population contribute equally to the model. In the mitochondria, the heteroplasmy level of a variant is a more important parameter than presence or absence for assessment of constraint. This is because there is selection to reduce the heteroplasmy level of deleterious variants in the mitochondria, reflecting that individuals can carry pathogenic variants and remain asymptomatic as long as the heteroplasmy level is below a ‘disease threshold’. For example, there are well-known pathogenic mtDNA variants that are present in gnomAD but only at lower heteroplasmy levels (e.g. m.3243A>G, maximum heteroplasmy value of 0.46 or 46% in gnomAD), and the median maximum heteroplasmy of heteroplasmies in gnomAD decreases as the predicted severity of the variant class increases¹⁰. Taking the same approach as the nuclear model would mean that deleterious variants only seen at lower heteroplasmy levels due to selection would contribute equally to benign variants present in gnomAD at high heteroplasmy (e.g. 100%), leading to underestimation of selection against deleterious variation. Therefore, we developed an approach for assessing observed:expected ratios that incorporated heteroplasmy (see section 5).

1.4 Why the nuclear method for calculating observed:expected ratio CI could not be applied

The nuclear constraint metrics computed for gnomAD provide 90% confidence intervals (CIs) around the observed:expected ratios, to capture the uncertainty around each estimate⁸. This is important because confidence in the observed:expected ratio can vary depending on sample (and

gene) size. The nuclear method for calculating the CIs is centered on a discrete Poisson distribution⁸. However, since we use continuous maximum heteroplasmy values to calculate observed:expected ratios (per above) instead of the number of unique variants assessed in nuclear models, we cannot use the discrete Poisson probability distribution. Therefore, we needed to develop an alternative approach for constructing ratio CIs that used a continuous probability distribution and could also appropriately model uncertainty around observed:expected ratios based on maximum heteroplasmy values (described in section 5.1).

1.5 Why the nuclear method for regional gene constraint could not be applied

The method used to identify regional constraint in nuclear genes is based on their exonic structure; applying a statistical framework to identify contiguous exonic segments that display a level of missense constraint that is significantly different from the gene's missense constraint. This method thus assumes exonic gene structure, testing all possible breaks between exons, but also that observed variant counts follow a Poisson distribution around the expected value. Neither of these assumptions are suitable here, since genes in the mtDNA do not have exons or introns and since maximum heteroplasmy cannot be modeled using a Poisson distribution (as a continuous variable). Therefore, we needed to develop an alternative method for identifying regional constraint in mtDNA (described in section 7). Nuclear methods also only assess regional missense constraint in protein genes; therefore, we also extended application to RNA genes in the mtDNA.

2. Mitochondrial composite likelihood model

We adapted the composite likelihood model described by Dietlein et al to quantify mutability in the human mtDNA¹¹. This model was developed for the study of passenger mutations in cancer nuclear exomes, which can have a sparsity of mutation counts per possible nucleotide context¹¹. This approach was therefore well suited to handle the possible sparsity of counts per context associated with the smaller genome size of the mtDNA. We made several adaptations to the composite likelihood model to enable its application to the mtDNA (see section 2.1). These modifications aimed to handle the known replicative strand bias for transitions⁵, unequal frequency of pyrimidines and purines in the reference sequence¹², as well as the inverted signature for transitions within the non-coding OriB-OriH region⁵.

We quantified mutability using de novo mutations ascertained from the literature and an in-house dataset (described below). Since the mutability of a mtDNA base can depend on the sequence context at the +1 and -1 positions^{5,13}, we quantified mutability in trinucleotide contexts. The composite model decomposes the mutational likelihood of each trinucleotide context into multiplicative factors, namely the effects of the base substitution type λ_t , the flanking 5' nucleotide λ_{-1} and the flanking 3' nucleotide λ_{+1} . The likelihood of the base substitution type λ_t can be further decomposed into the effect of the reference nucleotide $\lambda_{n(t)}$ and the mutation class $\lambda_{c(t)}$. A mutation likelihood score λ is obtained for every possible base substitution at every nucleotide in the mtDNA, where $\lambda > 1$ indicates increased likelihood of mutation, and $\lambda < 1$ a decreased likelihood compared with genome-wide average. Our mitochondrial composite likelihood model is described in detail below.

2.1 Why the nuclear composite likelihood model could not be applied

The composite likelihood model used to model mutation rates in the nuclear genome has several parameters and assumptions which are violated in the mtDNA. Firstly, the normalization coefficients are based on the approximately equal frequency of the four nucleotides (A, C, G, T) in the nuclear exome, which is not true in the mtDNA which has an unequal frequency of pyrimidines and purines in the reference sequence¹². This also has implications for the expected mutation counts assumptions applied. Secondly, this model assumes mutational strand symmetry, providing mutation likelihoods for pyrimidine base substitution types. This is not a suitable

assumption for the mtDNA, which has known replicative strand bias for transitions⁵. Thirdly, an inverted signature for transitions occurs within the non-coding OriB-OriH region⁵, necessitating special handling of this region within the overall composite likelihood model. Ultimately, multiple unique considerations required us to adapt the nuclear model for application to the mtDNA.

2.2 Mutation classification system and reference sequence

We classified the six pyrimidine base substitution types by whether the reference nucleotide C or T is on the reference or reverse complement strand (i.e. G and A in the reference strand). This classification of 12 base substitution types t and their reference nucleotide $n(t)$ is summarized in Supplementary Table 1. These mutations are further categorized into three mutational classes $c(t)$: transversions type I (class I), transversions type II (class II) and transitions (class III).

Supplementary Table 1: Mutation classification system used. This includes 12 mutation types, and three mutation classes. The pyrimidine mutation types are also shown.

Mutation	Reference nucleotide $n(t)$	Mutation type t	Mutation class $c(t)$	Pyrimidine mutation type $pyr(t)$	Pyrimidine reference nucleotide $pyr(n(t))$
C>A	C	1	I	C>A	C
C>G	C	2	II	C>G	C
C>T	C	3	III	C>T	C
T>A	T	4	I	T>A	T
T>C	T	5	III	T>C	T
T>G	T	6	II	T>G	T
G>T	G	7	I	C>A	C
G>C	G	8	II	C>G	C
G>A	G	9	III	C>T	C
A>T	A	10	I	T>A	T
A>G	A	11	III	T>C	T
A>C	A	12	II	T>G	T

We used the GRCh38 mitochondrial reference sequence NCBI accession NC_012920.1, also known as the revised Cambridge Reference Sequence (rCRS). Due to the different mutational signature within the non-coding OriB-OriH region spanning m.16197-16569 and m.1-191 (across

the artificial break)⁵, we computed mutability within the reference sequence excluding OriB-OriH (m.192-16196) separately to the OriB-OriH region.

2.3 Likelihood ratio for mutation at reference nucleotides

We counted the number of de novo mutations of each base substitution type t and summarized them into type count vector $\mathbf{v}^{type}$. Each element v_t^{type} corresponds to the number of base substitutions of type $t \in \{1, \dots, 12\}$. We use v_t^{type} to define the likelihood ratios for base substitution at the reference nucleotides:

$$\lambda_C = \frac{(v_1^{type} + v_2^{type} + v_3^{type})}{|v^{type}| \cdot P(n(t)=C)}$$

$$\lambda_T = \frac{(v_4^{type} + v_5^{type} + v_6^{type})}{|v^{type}| \cdot P(n(t)=T)}$$

$$\lambda_G = \frac{(v_7^{type} + v_8^{type} + v_9^{type})}{|v^{type}| \cdot P(n(t)=G)}$$

$$\lambda_A = \frac{(v_{10}^{type} + v_{11}^{type} + v_{12}^{type})}{|v^{type}| \cdot P(n(t)=A)}$$

where $|v^{type}| := \sum_t v_t^{type}$ (sum of all mutations across all 12 types), and $P(n(t))$ reflects our a priori assumption to observe mutations at each $n(t) \in \{A, C, G, T\}$ at the same frequency as the $n(t)$ within the reference sequence (i.e. the proportion of the $n(t)$ in the reference sequence).

2.4 Likelihood ratio for the mutation classes

We used v_t^{type} to define the likelihood ratios for the mutation classes I, II, III:

$$\lambda_I = \frac{(v_1^{type} + v_4^{type} + v_7^{type} + v_{10}^{type})}{|v^{type}| \cdot P(c(t)=I)}$$

$$\lambda_{II} = \frac{(v_2^{type} + v_6^{type} + v_8^{type} + v_{12}^{type})}{|v^{type}| \cdot P(c(t)=II)}$$

$$\lambda_{III} = \frac{(v_3^{type} + v_5^{type} + v_9^{type} + v_{11}^{type})}{|v^{type}| \cdot P(c(t)=III)}$$

where $|v^{type}| := \sum_t v_t^{type}$, and $P(c(t))$ is the probability of the mutation class, calculated as the sum of probabilities for each mutation type t in the class where

$$c(t) = \begin{cases} I & t = C > A, T > A, G > T, A > T \\ II & t = C > G, T > G, G > C, A > C \\ III & t = C > T, T > C, G > A, A > G \end{cases}$$

The probability of each type t is estimated using the relative frequency of nucleotides in the reference sequence. Specifically, we calculate this as the frequency of the reference nucleotide $n(t)$ multiplied by the frequency of the alternate nucleotide amongst all possible alternates in the reference sequence, i.e. $P(C > A) = P(n(t) = C) \cdot (P(n(t) = A)/(1 - P(n(t) = C)))$.

2.5 Likelihood of mutation type

We used the likelihood ratios $\lambda_{n(t)}$ for the reference nucleotide and $\lambda_{c(t)}$ for the mutation class to calculate the likelihood ratio of each base substitution type

$$\lambda_t = \lambda_{n(t)} \cdot \lambda_{c(t)}$$

for each $n(t)$ reference nucleotide and $c(t)$ mutation class defined in Supplementary Table 1.

2.6 Likelihood ratio for sequence context for class III mutations

We use v_t^{type} , $v_{t,p,n}^{seq}$ and $f_{n(t),p,n'}^{ref}$ (defined below) to determine the likelihood ratios for sequence context around class III mutations:

$$\lambda_{t,p,n} = \frac{v_{t,p,n}^{seq}}{v_t^{type} \cdot f_{n(t),p,n'}^{ref}}$$

where $v_{t,p,n}^{seq}$ is the count of nucleotide $n \in \{A, C, G, T\}$ at position $p \in [-1:1] \setminus \{0\}$ around base substitution of type $t \in \{C > T, T > C, G > A, A > G\}$, v_t^{type} is the count of base substitutions of type t , and $f_{n(t),p,n'}^{ref}$ is the frequency of nucleotide $n' \in \{A, C, G, T\}$ at position $p \in [-1:1] \setminus \{0\}$ around the reference nucleotide $n(t)$. Specifically, f^{ref} for class III mutations is calculated as:

$$f_{n(t),p,n'}^{ref} = \frac{f_{(n_p=n', n(t))}^{ref}}{P(n(t))}$$

where $P(n(t))$ is the proportion of the $n(t)$ in the reference sequence, and $f_{(n_p=n', n(t))}^{ref}$ is the frequency of observing nucleotide n' at position p around $n(t)$ across the total number of nucleotide sequences of length $(n_p, n(t))$ in the reference sequence.

2.7 Likelihood ratio for sequence context for class I and class II mutations

For class I and II mutations (transversions), due to their lower counts we calculate v^{seq} and $f_{n(t),p,n'}^{ref}$ by annotating them by their pyrimidine reference nucleotide (C or T) and calculating in a strand agnostic manner. While there is replicative strand bias for transitions, this has not been established for transversions^{5,14}, supporting this approach. Therefore, we classify the eight transversion mutation types $\mathbf{t}$ into their four pyrimidine mutation types $\mathbf{pyr}(\mathbf{t}) \in \{C > A, C > G, T > A, T > G\}$ (as per Supplementary Table 1).

We prepared $v_{\mathbf{pyr}(\mathbf{t})}^{type}$ where each element corresponds to the number of the $\mathbf{pyr}(\mathbf{t})$. We used $v_{\mathbf{pyr}(\mathbf{t})}^{type}$ as well as $v_{\mathbf{pyr}(\mathbf{t}),p,n}^{seq}$ and $f_{\mathbf{pyr}(n(\mathbf{t})),p,n'}^{ref}$ (defined below) to determine the following likelihood ratios for the sequence context around class I and II mutations:

$$\lambda_{t,p,n} = \frac{v_{\mathbf{pyr}(\mathbf{t}),p,n}^{seq}}{v_{\mathbf{pyr}(\mathbf{t})}^{type} \cdot f_{\mathbf{pyr}(n(\mathbf{t})),p,n'}^{ref}}$$

where $\mathbf{pyr}(\mathbf{t})$ is the pyrimidine mutation type of the mutation type $\mathbf{t}$, $v_{\mathbf{pyr}(\mathbf{t})}^{type}$ is the count of the $\mathbf{pyr}(\mathbf{t})$, and $v_{\mathbf{pyr}(\mathbf{t}),p,n}^{seq}$ is the count of nucleotide $n \in \{A, C, G, T\}$ at position $p \in [-1:1] \setminus \{0\}$ around $\mathbf{pyr}(\mathbf{t})$. For mutation types $\mathbf{t} = G > T, G > C, A > T, A > C$ where the reference nucleotide $n(\mathbf{t})$ is A or G, the complement $\bar{n}$ of the flanking nucleotide n at position $-p$ is counted instead. Therefore $v_{\mathbf{pyr}(\mathbf{t}),p,n}^{seq}$ is calculated as:

$$v_{\mathbf{pyr}(\mathbf{t}),p,n}^{seq} = \begin{cases} v_{\mathbf{pyr}(\mathbf{t}),p,n}^{seq} & \text{for } \mathbf{t} = C > A, C > G, T > A, T > G \text{ where } n(\mathbf{t}) = C, T \\ v_{\mathbf{pyr}(\mathbf{t}),-p,\bar{n}}^{seq} & \text{for } \mathbf{t} = G > T, G > C, A > T, A > C \text{ where } n(\mathbf{t}) = A, G \end{cases}$$

$f_{\mathbf{pyr}(n(\mathbf{t})),p,n'}^{ref}$ is the frequency of nucleotide $n' \in \{A, C, G, T\}$ at position $p \in [-1:1] \setminus \{0\}$ around the pyrimidine reference nucleotide $\mathbf{pyr}(n(\mathbf{t}))$, where:

$$\mathbf{pyr}(n(\mathbf{t})) = \begin{cases} C & n(\mathbf{t}) = C, G \\ T & n(\mathbf{t}) = T, A \end{cases}$$

For reference nucleotides $n(\mathbf{t})$ that are A or G, the frequency of the complement $\bar{n}'$ of the flanking nucleotide n at position $-p$ is counted instead. Specifically, $f_{\mathbf{pyr}(n(\mathbf{t})),p,n'}^{ref}$ is calculated as:

$$f_{pyr(n(t)),p,n'}^{ref} = \begin{cases} \frac{f_{(n_p=n',pyr(n(t)))}^{ref}}{P(pyr(n(t)))} & \text{where } n(t) = C, T \\ \frac{f_{(n_{-p}=\overline{n'},pyr(n(t)))}^{ref}}{P(pyr(n(t)))} & \text{where } n(t) = A, G \end{cases}$$

where $P(pyr(n(t)))$ is the proportion of the $pyr(n(t))$ in the reference sequence, and $f_{(n_p=n',pyr(n(t)))}^{ref}$ is the frequency of observing nucleotide n' at position p around $pyr(n(t))$ across the total number of nucleotide sequences of length $(n_p, pyr(n(t)))$ in reference sequence.

2.8 Composite likelihood of position mutability based on sequence context

We used the likelihood ratios for sequence context $\lambda_{t,p,n}$ to define the composite likelihood of the flanking nucleotide sequence on the mutation probability:

$$\lambda_{t,(n_{-1},\dots,n_1)} = \prod_{-1 \leq p \leq 1, p \neq 0} \lambda_{t,p,n_p}$$

where the effect of the flanking sequence $(n_{-1}, \dots, n_1) \in \{A, C, G, T\}^{1+1+1}$ of length $1+1+1$ on base substitutions of type t is modeled as a product of the effect of each flanking base, and $p = -1$ represents the 5' flanking nucleotide while $p = 1$ represents the 3' flanking nucleotide.

2.9 Composite likelihood of each mutation type at each position

We used the likelihood ratio for the mutation type and the composite likelihood of sequence context to determine mutability of every possible base substitution within the reference sequence, first excluding the OriB-OriH region. Specifically, we compute the mutability of each mutation class c at each mtDNA coordinate $m \in [192:16196]$ with reference nucleotide n_0 and flanking nucleotide n_p at position p as:

$$\lambda_{m,c} = \lambda_{n_0} \cdot \lambda_c \cdot \prod_{-1 \leq p \leq 1, p \neq 0} \lambda_{t,p,n_p}$$

The replicative strand bias for transitions has been shown to be inverted in a segment of the non-coding control region bounded by the OriB site at m.16197 and the OriH site at m.191, which we refer to as the 'OriB-OriH' region⁵. Therefore we quantified mutability in this region separately. For determining the likelihood of mutation type, we only use de novo mutations and reference

nucleotides within m.16197-191 to build $\mathbf{v}^{type}$, $\lambda_{n_0}^{ori}$ and λ_c^{ori} . For determining the likelihood of sequence context, we handle transitions and transversions separately. For transitions, we build $\mathbf{v}^{seq}$, $\mathbf{f}^{ref}$ and λ_{t,p,n_p}^{ori} using de novo mutations and reference nucleotides within m.16197-191. For transversions, due to their lower counts we apply the likelihood of sequence context calculated previously (λ_{t,p,n_p}) for m.192-16196, on the assumption that sequence context around transversions is similar within the OriB-OriH, given the distinct mutational signature reported is for transitions only. This can be summarized as the following for $m \in [1-191, 16196-16569]$:

$$\lambda_{m,c}^{ori} = \begin{cases} \lambda_{n_0}^{ori} \cdot \lambda_c^{ori} \cdot \prod_{-1 \leq p \leq 1, p \neq 0} \lambda_{t,p,n_p} & \text{where } c(t) = I, II \\ \lambda_{n_0}^{ori} \cdot \lambda_c^{ori} \cdot \prod_{-1 \leq p \leq 1, p \neq 0} \lambda_{t,p,n_p}^{ori} & \text{where } c(t) = III \end{cases}$$

3. Mitochondrial de novo mutation dataset

We aimed to estimate the de novo mutation rate for every possible base substitution in the mtDNA. Therefore, we used available datasets of mitochondrial de novo mutations to quantify mutability. We manually compiled a list of mitochondrial de novo mutations from the literature and supplemented these with an in-house dataset (Supplementary Table 2). This included germline de novo mutations, as well as somatic mutations (acquired in tissues and cancer) which have highly similar mutational mechanism and signature to germline mutations in the mtDNA^{5,15}.

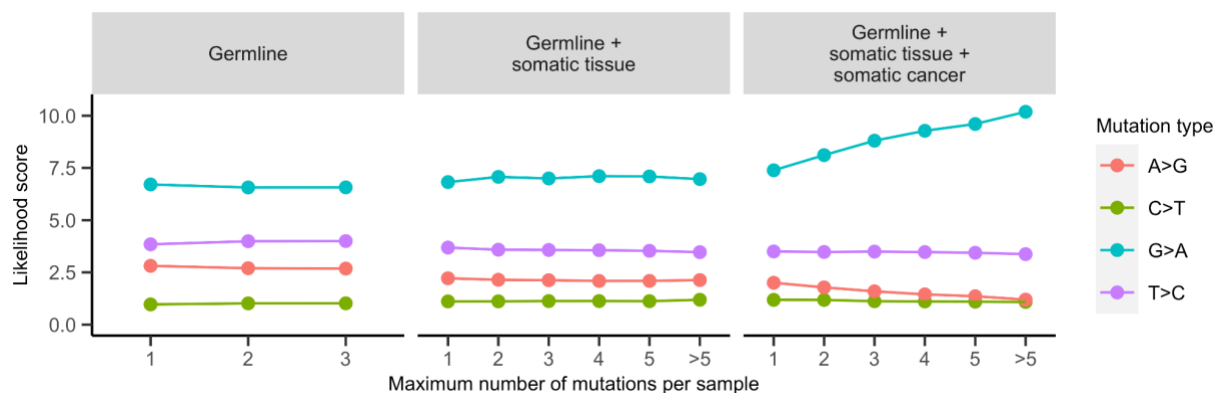
Supplementary Table 2: Overview of the de novo mutation datasets used. The reference, associated file used, and any other relevant information is provided.

Dataset	Source and reference	Notes
Dataset 1	Germline ¹³	Marked as de novo in Data S1.
Dataset 2	Germline and somatic tissue ¹⁶	Germline de novo is marked as ‘child’, and somatic as ‘somatic-gain’, in Table S3.
Dataset 3	Germline and somatic tissue ¹⁷	Table S3 (germline) and Table S4 (somatic tissue) was obtained from the Appendix file, and converted to txt file format. Dataset S1 was used to determine the major and minor allele at each position in both tables.
Dataset 4	Germline ¹⁸	De novos were manually identified from Dataset S1.
Dataset 5	Germline*	In-house dataset, described in section 3.1.
Dataset 6	Somatic tissue ¹⁵	Provided in Table S3. Inferred as aligned to the Yoruban reference sequence, so converted to the rCRS.
Dataset 7	Somatic cancer ¹⁹	In the file retrieved from https://ibl.mdanderson.org/tcma/download/TCMA-MutationSNV.tsv.zip

*This de novo dataset is unpublished.

We first performed checks to ascertain if there were any major differences between the de novo mutations identified from the various sources (germline, somatic tissue, and somatic cancer), as well as from samples with higher de novo counts. We compared the mutation likelihoods produced

from each source of de novo mutations, for transitions. This revealed that while the likelihoods estimated by germline and somatic tissue datasets were similar, the likelihoods estimated by somatic cancer mutations for G>A (and to a lesser extent A>G) mutations diverged as the total number of de novo mutations per sample increased (Supplementary Fig. 1). Therefore, we decided to only use somatic cancer mutations that were from samples with only one de novo mutation, since these were similar to germline and somatic tissue datasets, and given we ultimately aimed to estimate the germline de novo rate. This resulted in a final list of 4216 de novo mutation allele counts, including 3799 within the reference sequence excluding the OriB-OriH region (m.192-16196, including 3421 transitions and 378 transversions), and 417 within the OriB-OriH region (m.16197-191, including 385 transitions and 32 transversions).



Supplementary Figure 1: Mutation likelihoods of de novo transitions by source. Data represents samples from germline ($n=522$), somatic tissue ($n=1308$), and somatic cancer ($n=2170$) datasets. The likelihood of G>A mutations increases linearly as the sample de novo mutation count increases.

3.1 In-house de novo mutation dataset

De novo mtDNA mutations identified from Simons Foundation Powering Autism Research for Knowledge (SPARK)²⁰ study participants were included in the de novo dataset used for mutational model building, obtained as follows. Whole genome sequencing data from SPARK releases 1-3 were retrieved from SFARI Base for probands and their family members. These data are subject to quality control prior to release, including relatedness and contamination checks. Only unaffected siblings with maternal data were used to generate de novo calls, since the mtDNA has been

implicated in autism pathogenesis²¹. This left 1690 mother-child pairs for analysis. GATK Mutect2 “mitochondria-mode” was used to call variants, as described previously¹⁰ with the exception that filters “possible_numt” and “mt_many_low_hets” were applied. All PASS variants in each child at >1% heteroplasmy level that were not a PASS variant with >1% heteroplasmy in their mother were regarded as candidate de novo variants. Filtering was then applied to produce a high-confidence list by excluding any candidate de novos where (1) a variant was called at same position in the mother, including non-PASS calls, or (2) if the candidate de novo was a SNV and an indel was called at the same or an adjacent position (+1 and -1 bases) in the child (including non-PASS indels). From this 540 de novo mutations were identified, a de novo rate consistent with that reported by Wei et al¹³, which were used in the de novo dataset. We then evaluated this cohort to test our assumption in mutational model building that each de novo mutation occurred in the reference trinucleotide context. This showed that 537 of the 540 (>99%) de novos occurred in the reference trinucleotide, in line with our assumption. Only 3 de novos occurred at a site with a variant at a flanking nucleotide, which were all homoplasmic haplogroup variants, indicating that background variation had minimal impact on mutation model building.

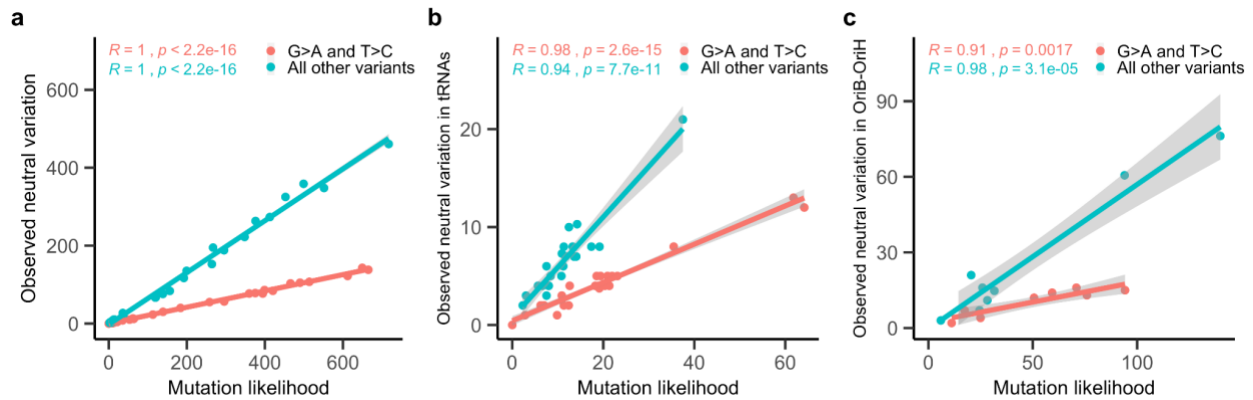
4. Validation of mutation model

The observed level of neutral variation in a population dataset should correlate with mutation rates, since it is not subject to strong selection. Therefore, we measured the correlation between the observed level of neutral variation in gnomAD and their mutation likelihoods, to confirm the predictive value of our mutational model. We used haplogroup variants from PhyloTree Build 17²² to represent neutral variation given they are passed down a maternal lineage at homoplasmy, supplemented with variants at non-conserved sites to increase representation of different mutation types (i.e. transversions), defined as being in the lowest decile of PhyloP scores (equivalent to a score of <-7.25 for reference excluding OriB-OriH and <-2.78 for the latter). Therefore, any SNV in PhyloTree (a database of mitochondrial haplogroup variants) or at a site in the lowest PhyloP decile (defined above) were used (n=8284 unique variants, Supplementary Dataset 8). This approach enabled us to include neutral variation in protein and RNA genes, and in non-coding loci.

We fit the highly mutable G>A and T>C variants separately, akin to how CpG transitions are handled separately in nuclear models⁸, given prior analyses had shown that these mutation types were likely saturated in gnomAD¹⁰. Neutral variation was assessed in all mtDNA genes as well as in the non-coding sequence excluding OriB-OriH. The observed sum maximum heteroplasmy of the neutral variants in each gene or locus in gnomAD was calculated and regressed against their sum mutational likelihood; no variant frequency or population filters were applied for this step. We observed that the mutation likelihoods were highly correlated with the observed level of neutral variation in each locus in gnomAD, with Pearson correlation coefficients of $R>0.99$ (Supplementary Fig. 2a, p-value $<2.2e-16$), which remained high when restricting to the smaller tRNA genes (Supplementary Fig. 2b). A comparison of correlations using the *cocor* R package²³ showed that the correlation between mutation likelihoods and observed levels ($R=1$) was significantly stronger than that between mutation likelihoods and locus length ($R=0.93$, *cocor* p-values $<1 \times 10^{-03}$). Multiple regression analysis confirmed that the association between the mutation likelihoods and observed variation ($p<2.2e-16$ for G>A/T>C and all other mutations) was much more significant than that between locus length and observed levels ($p=0.03$ and $p=0.134$ for G>A/T>C and all other mutations respectively). We validated the non-coding OriB-OriH region separately to the rest of the reference sequence. To enable assessment of correlation in this region, we segmented it into eight non-overlapping blocks of approximately equal size, of similar length

to the tRNA genes (70 bp). We observed a high correlation between mutation likelihoods and observed neutral variation within the OriB-OriH region ($R>0.9$, $p<3.1e-05$ and $p<0.002$) (Supplementary Fig. 2c). These data establish the predictive value of the mutational model.

Previously we showed that nearly all possible synonymous for the most mutable base changes are observed in gnomAD at homoplasmy¹⁰, supporting that background selection is not strong enough to prevent a neutral variant from reaching homoplasmy in a population sample the size of gnomAD. This supports that background selection likely has minimal impact on the population maximum heteroplasmy of neutral variants, which is the metric we use to assess constraint. This is also in line with the effects of background selection on maximum heteroplasmy being minimized once a neutral variant reaches fixation (i.e. homoplasmy) in a maternal lineage, since such variants can remain at homoplasmy despite selection against new deleterious alleles.



Supplementary Figure 2: Assessing the predictive value of the mutational models. (a-c) For each locus ($n=39$ including control region excluding OriB-OriH, and all other non-coding loci) (a), tRNA gene ($n=22$) (b) or equally sized blocks in the OriB-OriH region ($n=8$) (c) the observed sum maximum heteroplasmy of neutral variants is plotted against the sum of their mutation likelihood scores. The highly mutable G>A and T>C were fit separately (shown in red). Pearson correlation coefficient (R) and its p -value (p) computed using a two-sided test is shown. The line represents the linear model fit, and the gray shading the 95% confidence level around it.

5. Assessment of mitochondrial constraint

The large size and diversity of gnomAD, its inclusion of homoplasmic and heteroplasmic variation, depletion of severe pediatric disease, and stringent quality control makes it well-suited for the study of constraint⁸. We assessed mitochondrial constraint in gnomAD as a ratio of observed to expected variation. Nuclear models of constraint applied to gnomAD assess the number of unique variants within a gene or region in the population⁸. Since selection can occur against both the number of variants and their heteroplasmy level in the mitochondria¹⁰, we assess the observed and expected sum maximum heteroplasmy of mtDNA variants in the population¹⁰. This value represents the maximum heteroplasmy (fraction of alternate to total reads, range 0.0-1.0) that the variant is observed at across all individuals in gnomAD. Simulation of germline mtDNA mutation and heteroplasmy drift across generations support the validity of using the mutational model to estimate the expected sum maximum heteroplasmy (see below). Each mtDNA variant in gnomAD v3.1 has a maximum heteroplasmy between 0.1-1.0 since only variants with heteroplasmy >10% were included in this release; variants not observed are assigned a maximum heteroplasmy of 0.

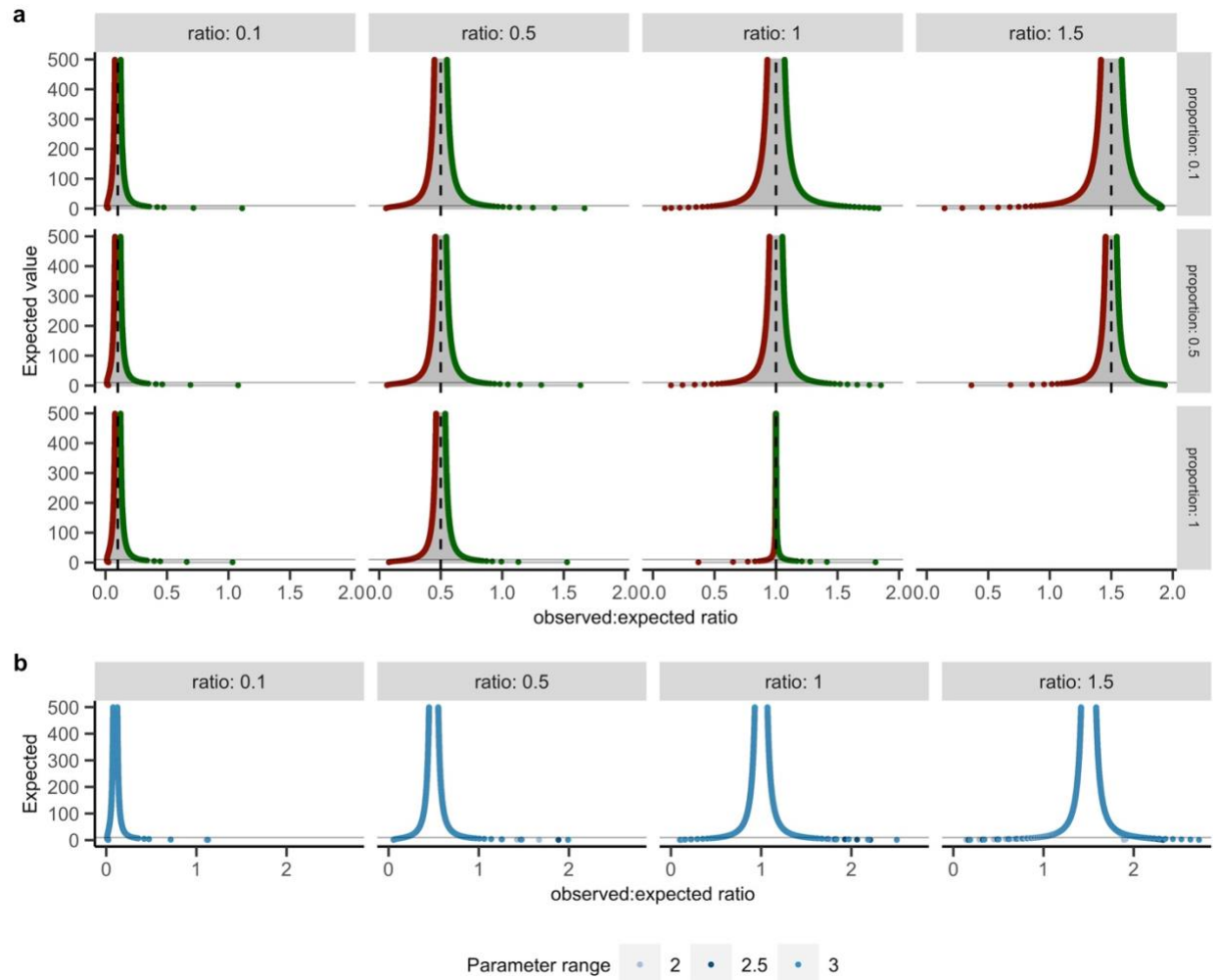
The observed value for a variant class and or within a gene or region is determined by summing the maximum heteroplasmy of every possible single nucleotide variant (SNV) belonging to this group (e.g. all missense within a protein-coding gene). We calculated the expected sum maximum heteroplasmy by summing the mutation likelihoods of every possible variant within the variant class/region of interest, and then applying the linear equations fit on mutation likelihoods and observed sum maximum heteroplasmy of neutral variation (ascertained using Phylotree and PhyloP scores as per section 4) in gnomAD; the highly mutable G>A and T>C variants were fit separately, as were variants within the non-coding OriB-OriH region. Variants at six artifact prone sites not called in gnomAD were excluded from observed and expected value calculations (m.301, m.302, m.310, m.316, m.3107, and m.16182)¹⁰. We applied a clamp to bound the expected value by biological limits, such that it cannot be less than 0 or greater than the maximum possible value. This method can currently only be applied for the assessment of SNVs, and therefore provides the observed:expected ratio for SNVs specifically.

Nuclear constraint models include corrections for low coverage regions and methylation levels⁸. We note that we do not apply these given the high and even coverage of the mtDNA in gnomAD¹⁰, and lack of robust data on mtDNA methylation^{24,25}.

5.1 Confidence interval and OEUF

We provide a 90% confidence interval (CI) around the observed:expected (oe) ratio, adapting the method used for nuclear constraint models applied to gnomAD⁸. For a given pair of observed and expected values, we compute the density of a beta distribution where x =proportion possible observed, and parameters a =expected value times a varying parameter (ranging between 0-2.5, at intervals of 0.001 for a total 2500 parameters) + 1, and b =maximum possible observed - a . This varying parameter range is sufficient to sample the distribution of ratios for confidence interval construction (Supplementary Fig. 3). For observed sum maximum heteroplasmy values of 0, we used 0.1 (the minimum possible in gnomAD) to enable assessment, since beta distributions cannot use 0 values. The cumulative density function of this function was computed and the value of the varying parameter extracted at points corresponding to 5% and 95% to indicate the bounds of the 90% CI around the oe ratio. A very wide CI with low utility is produced for expected values <10 (Supplementary Fig. 3), and these are thus excluded from most analyses. This CI provides a probabilistic statement about the observed:expected ratio, such that it represents a range of plausible values, and is not a definitive statement about the ratio value.

The oe ratio 90% CI upper bound fraction (OEUF) provides a conservative measure of constraint and is used for most analyses. Because our confidence in the ratio can vary depending on sample size, the OEUF is useful to capture the uncertainty around an estimate. Genes or variant classes with higher OEUF values will include those where the sample size is too small to confidently identify constraint as well as those not constrained, while low OEUF values should be confidently depleted for variation. This makes the OEUF a more conservative measure than the ratio.



Supplementary Figure 3: Confidence interval characteristics. (a) The observed:expected ratio 90% confidence interval (CI) width is shown in gray across expected values 0-500 for different ratios (0.1, 0.5, 1.0, 1.5) and proportions of possible variation (0.1, 0.5, 1.0, based on expected value), for varying parameter range 0-2.5. The CI lower bound is in maroon, and the upper bound fraction (OEUF) in green. Dashed vertical lines represent the observed:expected ratio. When proportion is 1.0 the ratio cannot be >1 ; hence no values are shown for ratio 1.5. (b) The CI is shown for different ratios (0.1, 0.5, 1.0), colored by the maximum value in the varying parameter range used. Data for proportion 0.1 is shown. For (a-b) a horizontal line is shown at expected=10.

6. Simulation of germline mtDNA mutation and heteroplasmy

To support the utility of using the mutational model to assess maximum heteroplasmy, we adapted a computational framework developed by Colnaghi et al²⁶ to simulate germline mtDNA mutation and heteroplasmy drift across generations, and used it to validate a correlation between mutation rates and maximum heteroplasmy for neutral mutations. We applied the parameters drawn from human data by Colnaghi et al, and adapted their approach of modeling mtDNA replication and cell division during oogenesis (development of the female germline) as follows:

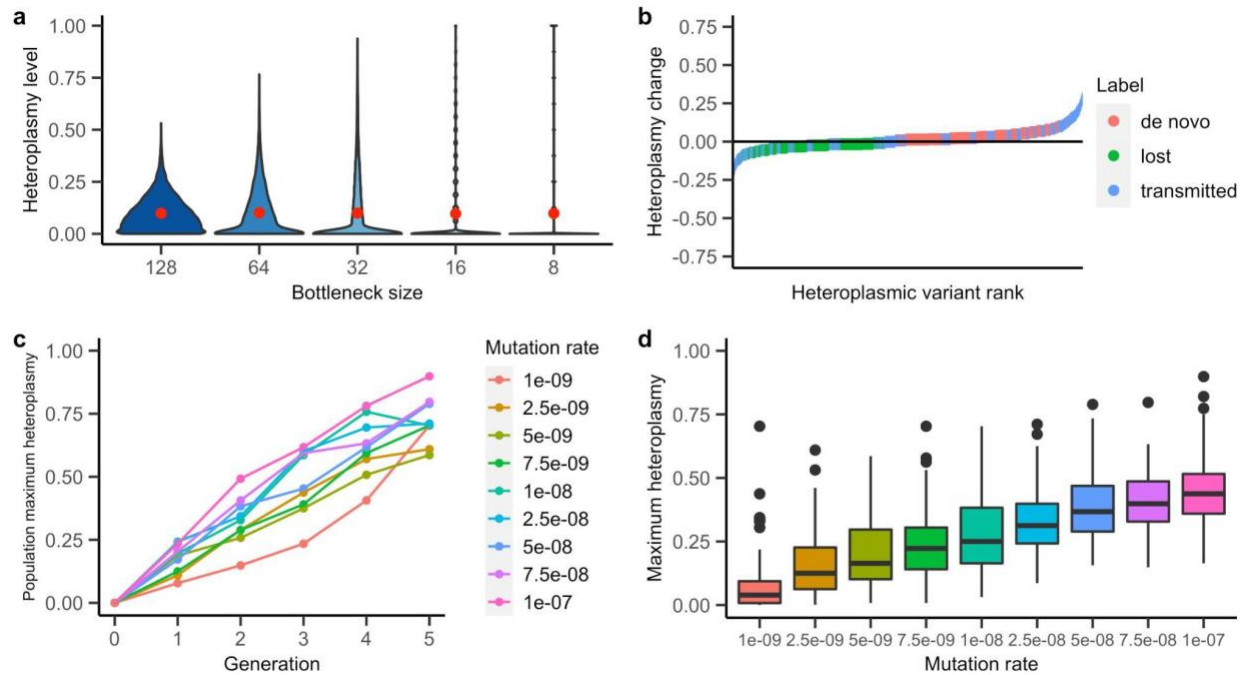
(1) Starting from a zygote with 2^{19} mtDNA copies (approximately half a million), we first simulated the formation of a primordial germ cell (PGC) through a series of 12 cell divisions without mtDNA replication per Colnaghi et al, with random partition of mtDNA to the daughter cells at each division drawn using a hypergeometric distribution. This results in the formation of a bottleneck, such that the PGC has 128 mtDNA copies (or fewer as specified).

(2) We then modeled the formation of a primary oocyte from the PGC via a series of 18 cell divisions with mtDNA replication per Colnaghi et al, during which constant mtDNA copy number is maintained. This involves first modeling the probability of a mutation during replication, achieved via a binomial distribution for 10 mutation rates between 10^{-9} - 10^{-7} per base pair, and then simulating cell division as above. We also included the possibility of back mutation during replication at the same mutation rate and assumed that the PGC survives the random cell death that occurs during this process.

(3) We then simulated the formation of a mature oocyte from the primary oocyte, via clonal amplification of the mtDNA without cell division until the starting mtDNA copy number of 2^{19} is reached. The probability of mutation during replication is modeled as above. We extend this framework across multiple generations by assuming that the heteroplasmy level in the mature oocyte is equal to the heteroplasmy level in the zygote it forms, providing the starting heteroplasmy level for the next generation.

This process from zygote to mature oocyte was repeated for five generations, for 10,000 maternal lineages. This method has a few simplifying assumptions; we follow the heteroplasmy of only one

mutation per lineage, we assume that it is neutral and not subject to selection, and we assume one pool of mtDNA per cell such that we don't consider multiple segregating units. The validity of our adaptation was supported by reproducing analyses reported by Colnaghi *et al*²⁶ and Wei *et al*¹³ (Supplementary Fig. 4a-b). Simulation of mtDNA mutation and drift across generations supported that mutation rates correlate with maximum heteroplasmy under neutrality (Supplementary Fig. 4c-d), in line with reports that neutral mutations drifting to homoplasmy (i.e. to maximum heteroplasmy of 1.0) increases linearly with mutation rate⁵, corroborating our approach.



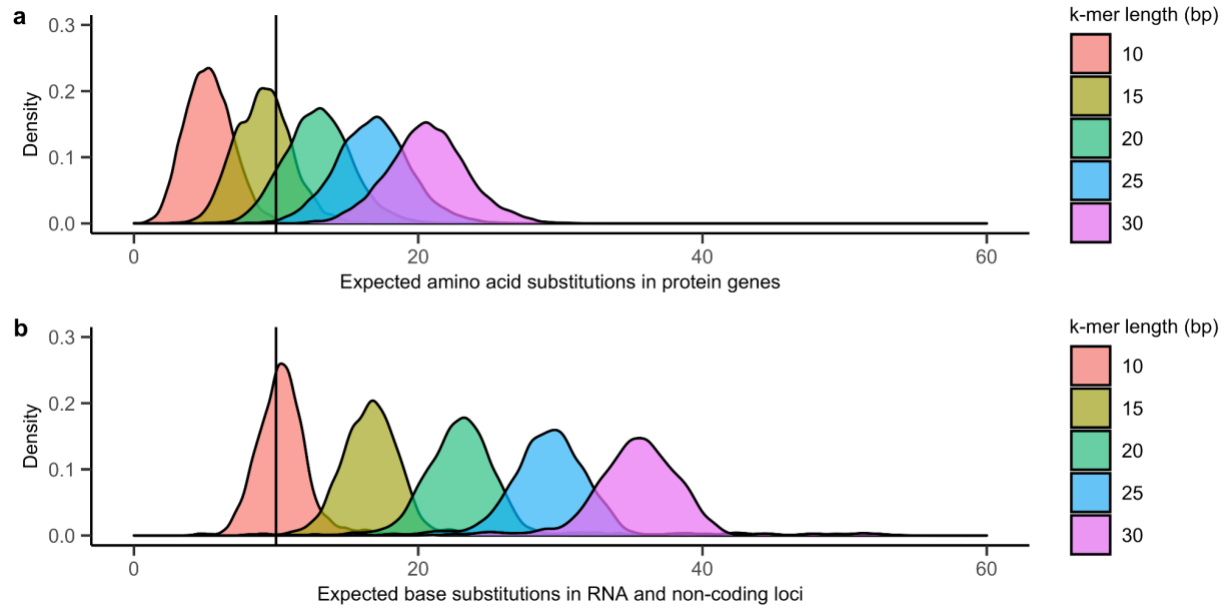
Supplementary Figure 4: Simulation of germline mutation and heteroplasmy in maternal lineages. (a) The simulated heteroplasmy distribution of mutations in $n=10,000$ offspring who inherited a mutation with heteroplasmy 0.1, for five bottleneck sizes; the mean is shown in red. This plot reproduces an analysis by Colnaghi *et al*²⁶, using a mutation rate of 10^{-8} . For (b-d) we used their standard bottleneck size of 128. (b) Heteroplasmy changes between offspring (generation 5) and mothers (generation 4) across $n=10,000$ lineages, ordered by degree of shift. *De novo* represents variants with heteroplasmy >0.01 in offspring and <0.01 in mothers, lost have heteroplasmy <0.01 in offspring and >0.01 in mothers, and transmitted have heteroplasmy >0.01 in offspring and mothers, to match criteria used by Wei *et al*¹³. A mutation rate of 10^{-8} was used; starting heteroplasmy was 0 for generation 0. (c) Maximum heteroplasmy across $n=10,000$

lineages across five generations, colored by mutation rate. **(d)** Maximum heteroplasmy distribution from $n=1000$ bootstrap replicates sampling $n=100$ individuals from the population ($n=10,000$) at generation five; starting heteroplasmy was 0 for generation 0. Boxplot elements include: center line, median; box limits, 25th and 75th percentiles; minima and maxima, 1.5x interquartile range; points, outliers.

7. Regional constraint

Although motivated by analyses of nuclear regional missense constraint²⁷, these methods were not applicable to the mtDNA, in part due to utilizing exon boundaries. Therefore we developed a method that compares the missense observed:expected (oe) ratio of all possible regions within a gene to the gene itself, to identify any regions which have a missense oe ratio that is significantly lower than the gene, an approach that is computationally feasible within the mtDNA. We then adapted the approach taken by Davydov et al²⁸ to reduce these data to discrete intervals of regional missense constraint by applying a greedy algorithm, and calculated the false discovery rate of each using random permutations of the reference sequence. The final method applied is as follows.

First, a list of all possible regions within each protein gene (termed *k*-mers) was compiled, of lengths ranging from a user-defined minimum to the maximum possible (gene length - 1). These *k*-mers spanned codons and *k*-mers 10 codons or larger were assessed (i.e. p.1-10); this minimum was used to enable all *k*-mers to have expected value >10 (Supplementary Fig. 5a). Starting from the first codon position, a *k*-mer of the minimum length is drawn (e.g. p.1-10) and the missense oe ratio within it is calculated (where missense is the most severe consequence). The *k*-mer length is then increased by one codon (e.g. p.1-11), and the missense oe ratio within this *k*-mer is calculated. This process is repeated until all possible lengths with that *k*-mer start position are evaluated (e.g. p.1-12, p.1-13, etc), after which the *k*-mer start position is moved by one and the process is repeated (e.g. p.2-12, p.2-13, etc). For each *k*-mer, the probability of observing a missense oe ratio that is less than or equal to the gene's missense oe ratio is calculated by computing the cumulative density function of a beta distribution with parameters $a = (\text{expected value} * \text{gene oe ratio}) + 1$ and $b = \text{maximum possible value} - a$. All *k*-mers with p-value >0.01 are discarded; we also discard any remaining *k*-mers with a missense oe ratio 90% CI that overlaps the gene's CI. A greedy algorithm is then applied, such that any *k*-mer which overlaps another *k*-mer with a lower (more significant) p-value is discarded; for overlapping *k*-mers with the same p-value the longest *k*-mer is retained. This produces a list of non-overlapping *k*-mers that are significantly more missense constrained than the gene, which we term candidate areas of regional missense constraint (RMC). This process is summarized in Supplementary Fig. 6.

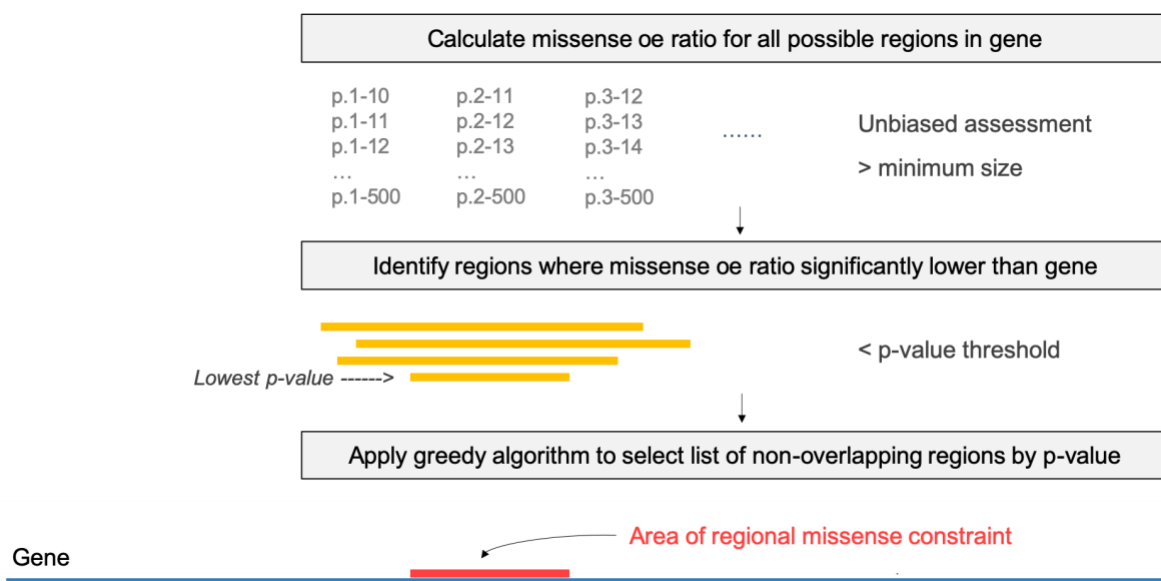


Supplementary Figure 5: Distribution of expected values for *k*-mer by length. (a-b) Density plots showing the distribution of expected values across all possible *k*-mers of each length (in bp) for missense variants in protein genes (10 bp, $n=11284$; 15bp, $n=11248$; 20 bp, $n=11208$, 25 bp, $n=11168$, 30 bp, $n=11124$) **(a)** or base substitutions within non-protein loci (10 bp, $n=5147$; 15bp, $n=5102$; 20 bp, $n=5057$, 25 bp, $n=5012$, 30 bp, $n=4967$) **(b)**. A vertical line marking an expected value of 10 is shown.

We calculated the false discovery rate (FDR) for each candidate area of RMC by assessing random permutations of the reference sequence. Specifically, we generated 1000 permutations of each gene by randomly shuffling the order of the codons, and then applied the process described above to yield a set of non-overlapping significant *k*-mers (if any) which we regard as false positives. We note that the entire gene sequence is permuted, such that any candidate areas of RMC are included, which may contribute false positives. We calculated the FDR of each candidate RMC area by counting how many permutations of the same gene had a false positive result of the same length and $\leq$ p-value to the candidate, and divide this number by the total number of permutations ($n=1000$). We then filtered out any candidate RMC areas with $FDR > 0.1$. Longer candidates are more likely to have lower FDR, and therefore a filtered candidate may overlap another significant *k*-mer that has a $FDR < 0.1$ but was discarded by the greedy algorithm (i.e. a longer region discarded due to having a higher p-value than the filtered candidate). To recover these, we reapply the greedy

algorithm after removing regions with $FDR > 0.1$ from both the real and permuted sequence, and recalculate the FDR for each resultant candidate RMC. The process is repeated if needed until all remaining candidates satisfy the $FDR < 0.1$ threshold, or no candidates remain. We note that only 9 of 39 initial candidate RMC areas did not pass the FDR threshold. This process yields a list of high-confidence RMC areas that are used for all analyses (Supplementary Dataset 2).

Regional constraint in the rRNA genes was calculated using the same process with the following modifications: (1) k -mers are measured in bp, (2) all possible base substitutions (i.e. SNVs) are included in observed and expected calculations, (3) k -mers 20 bp or larger were assessed (to enable all to have expected value > 10 per Supplementary Fig. 5b), and (4) the random permutations of each gene are achieved by shuffling the order of the bases.



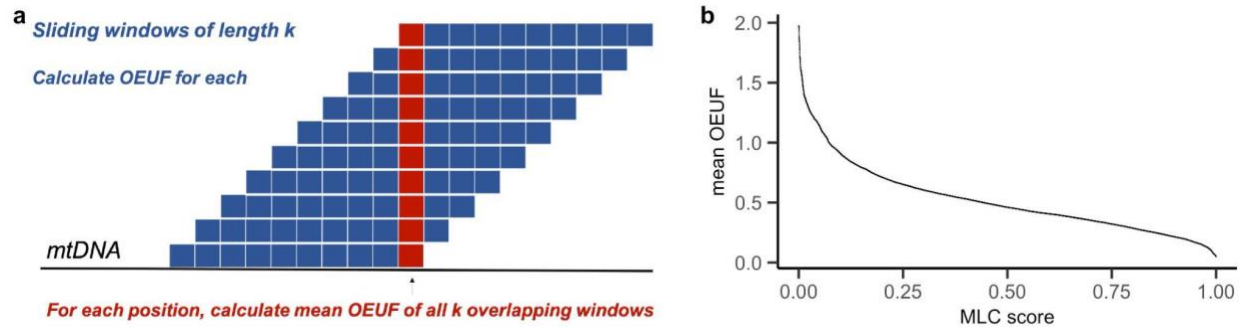
Supplementary Figure 6: Overview of the regional missense constraint method. The missense observed:expected (oe) ratio of all possible regions within a gene are assessed to identify any that have a missense oe ratio that is significantly lower than the gene's. A greedy algorithm was then applied to reduce these data to discrete intervals of regional missense constraint by p-value. The same method is applied to random permutations of each gene sequence for estimation of the false discovery rate.

8. Local constraint score

We were interested to identify the most constrained sites in the entire human mtDNA. To do this, we employed a sliding window approach for capturing the strength of selection against base or amino acid substitutions at and around each base position, which we term local constraint. From an initial start position of m.1, a window of length k is drawn and the observed:expected (oe) ratio and 90% confidence interval (CI) of substitutions within the window is calculated. The window start position is then moved by 1 bp, and the oe ratio and CI of the next window of length k is calculated. This process is repeated until all possible overlapping windows of length k in the mtDNA are evaluated; windows which overlap two different loci are therefore included. For positions in non-coding and RNA loci all possible base substitutions (i.e. SNVs) are included in calculations, while for positions in protein genes only missense variants are included to restrict assessment to amino acid substitutions (where missense is the most severe consequence). We used a uniform window length k of 30 bp, which for proteins captures 10 amino acids; this length was used to enable all k -mers to have an expected value >10 (Supplementary Fig. 5).

The local constraint of each base position in the mtDNA was then measured by taking the mean oe ratio CI upper bound fraction (OEUF) of all k windows that overlap the position. This approach therefore incorporates data from neighboring bases, such that bases in the closest proximity will contribute more of the signal than bases further away (Supplementary Fig. 7a). These mean OEUF values were then percentile ranked, such that each position was assigned a score between 0 and 1, which we term the mtDNA local constraint (MLC) score (Supplementary Fig. 7b). The higher the score, the more locally constrained the position is. The score for every base position is provided in Supplementary Dataset 6. Since this score is a measure of local intolerance to missense variants specifically in proteins, we also assigned scores to non-missense variants in protein genes to enable application to all possible SNVs. The MLC scores for variants were obtained as follows: non-coding, RNA and missense variants were assigned their positional score, synonymous variants were assigned a score of 0, stop gain were assigned a score of 1, and start and stop lost variants were assigned a score of 0.70. These assigned values for non-missense variants in protein genes are based on the OEUF of these functional variant classes (synonymous OEUF=1.01, stop gain OEUF=0.015, start lost OEUF=0.35, stop lost OEUF=0.34) and their placement within the

spectrum of mean OEUF values (Supplementary Fig. 7b). MLC scores for every possible mtDNA SNV are provided in Supplementary Dataset 7.



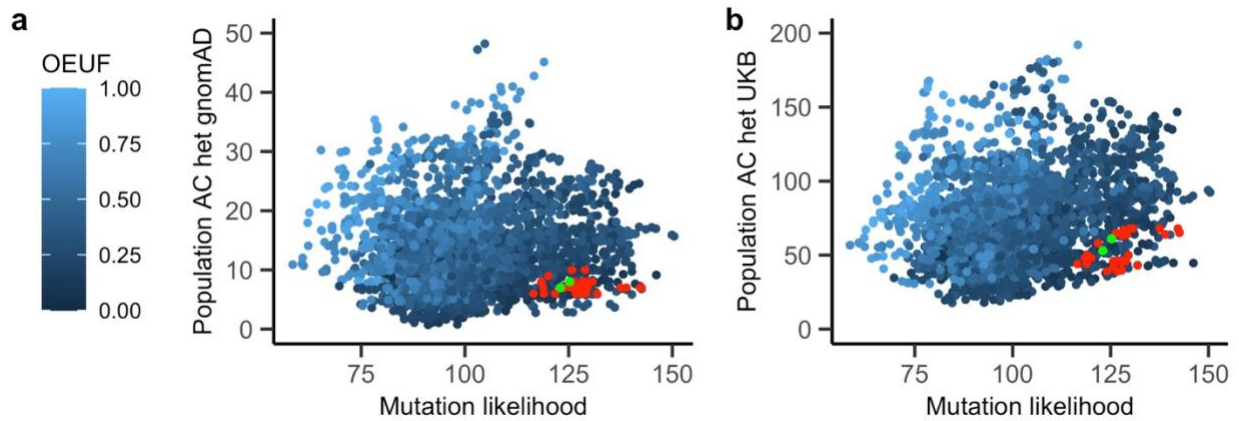
Supplementary Figure 7: Overview of the mtDNA local constraint (MLC) score. (a) Schematic overview of the method used to generate the MLC score. A sliding window of uniform length k is moved across the mtDNA at +1 base pair intervals. The observed:expected (oe) ratio of amino acid or base substitutions, and its 90% confidence interval (CI), was calculated for each window. For positions in protein genes only missense variants were included in calculations, while for all other positions all base substitutions are used. The mean oe ratio CI upper bound fraction (OEUF) of all k overlapping windows is computed for each position, and then percentile ranked to produce the MLC score. Note non-missense variants in protein genes are handled separately. **(b)** Relationship between the mean OEUF and MLC score ($n=16568$ positions).

9. Comparison of constraint with traditional methods

The development of a constraint model for the mitochondrial genome advances beyond traditional approaches that have been historically used to analyze the mtDNA; population allele frequency and conservation. Constraint modeling has key advantages over these traditional approaches. For example, one challenge with using allele frequencies to distinguish variants involved in disease is that most variants are very rare; ~90% of possible mtDNA SNVs have population frequency <0.0001 in gnomAD. This is a key consideration of using allele frequencies; i.e. across the many variants with low frequency, how to distinguish between those that are rare due to low mutation rate vs those that are rare due to selection reducing their frequency (i.e. deleterious)? This is what constraint models aim to do, providing a tool to help distinguish these scenarios.

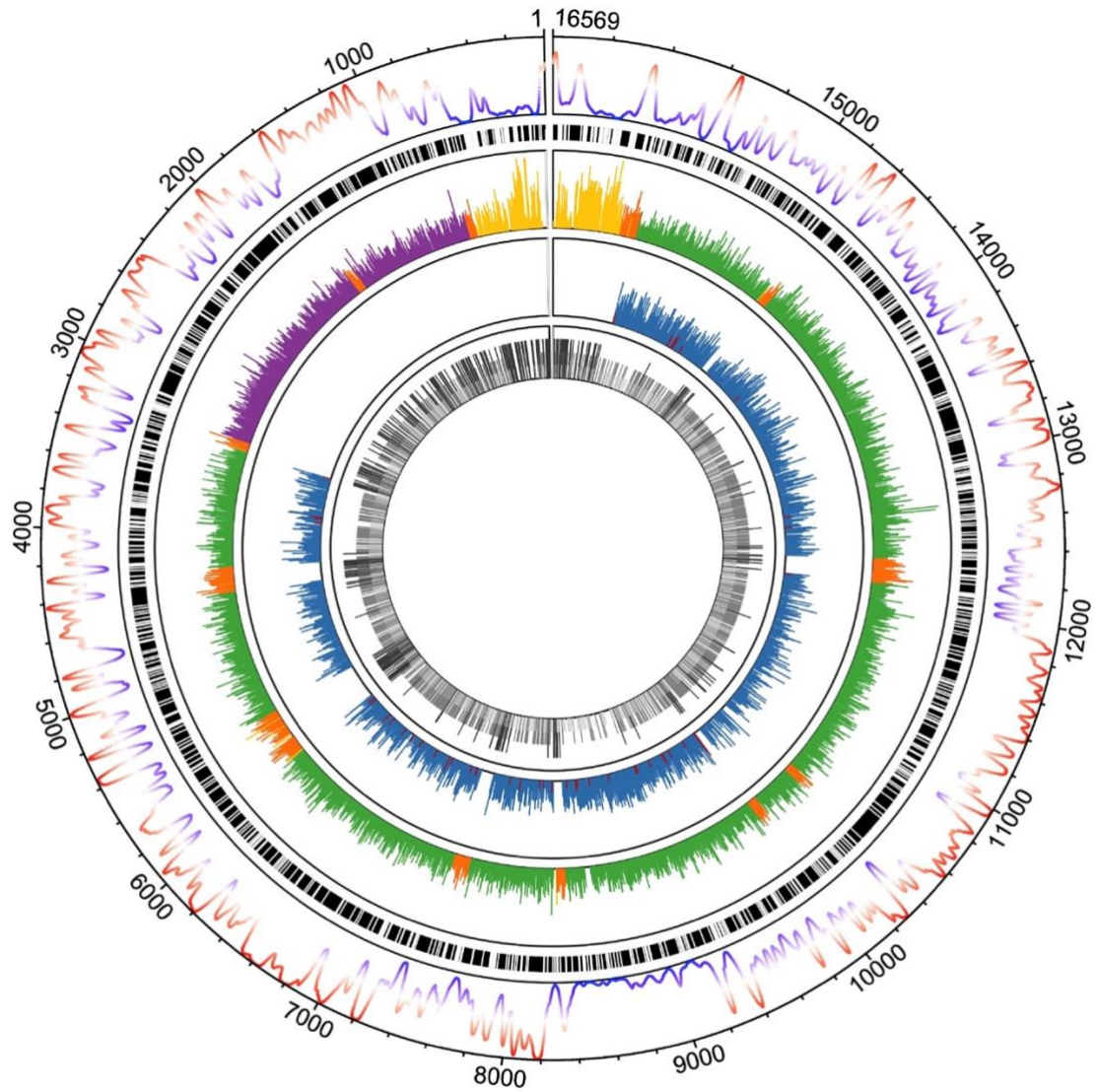
To illustrate this, Supplementary Figure 8 shows that regions with the same population frequency (y-axis) have different mutability (x-axis) for heteroplasmies; and how incorporating this information enables recognition of sites depleted of variation due to selection (i.e. constrained, dark blue color) vs low mutability (light blue). In this example, we highlight regions with the functional modification 2'-O-methylation (red color) and those spanning sites in the functional study (green), which were among the most constrained sites. Simply looking at allele frequencies would rank comparably to non-constrained sites with similar frequency due to lower mutability.

A traditional approach of simply looking at population frequency is also unable to effectively answer questions such as ‘what is the most constrained site in mtDNA?’ because a large proportion of bases are invariant in population databases. Constraint is a tool to help distinguish which sites are depleted of variation due to selection vs low mutability. For example, a comparison of MLC scores with frequency data did not reveal any striking similarities, where heteroplasmies with the same frequency in UK Biobank have a spectrum of MLC scores (Supplementary Fig. S9). Accordingly, there was a low correlation between MLC and population frequency for these variants in UK Biobank, and also between MLC and phyloP scores (Extended Data Fig. 10). The latter may reflect that constraint enables better capture of selective pressures in humans; warranted given metabolic requirements vary widely across species. Overall, these results emphasize that constraint modeling enables novel insights over traditional methods.

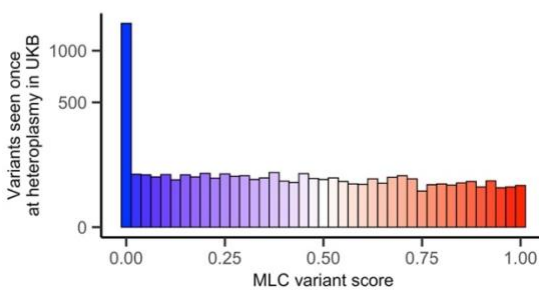


Supplementary Figure 8: Regions with the same population allele frequency have different mutability. (a) The plot shows population allele count (AC) total for heteroplasmies in gnomAD vs total mutational likelihood for equally sized regions in the RNA genes, which have the same number of possible variants. Regions are colored by their observed:expected ratio CI upper fraction (OEUF) value as a measure of constraint; those in red encode highly constrained sites with functional modifications (2'-O-methylation) while those in green span highly constrained sites included in the functional study (Figure 5). Regions are overlapping 30-mer windows ($n=2989$), as used to compute MLC. The y-axis is zoomed to total AC between 0-50. **(b)** The plot shows population heteroplasmic AC total in the UK Biobank (UKB) vs total mutational likelihood for the same equally sized regions ($n=2989$) in (a). Regions are colored using their OEUF values from (a) and the y-axis is zoomed to total AC 0-200. UKB variant data in (b) is per Hong, Battle, *et al*³.

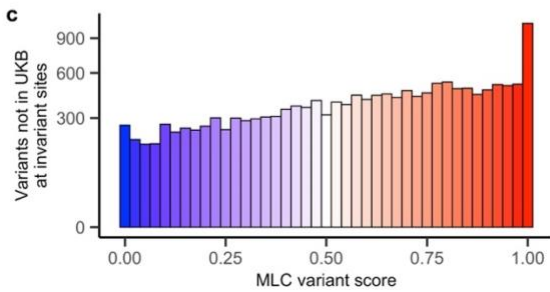
a



b



c



Supplementary Figure 9: MLC scores vary across heteroplasmies with the same allele frequency. (a) The circos plot tracks from outer (i) to inner (v) display: (i) MLC scores, ranging from 0.0 (dark blue) to 1.0 (dark red); (ii) bases conserved in chimpanzees are colored in black,

while those non-conserved and encoding base or amino acid substitutions are white; (iii) all synonymous (green), tRNA (orange), rRNA (purple), and non-coding (yellow) heteroplasmic sites; (iv) all missense (blue) and stop gain (red) heteroplasmic sites; and (v) positions with no heteroplasmy, where three or more adjacent invariant positions are colored light gray (lowest height), two adjacent positions are medium gray (medium height), and singlets are dark gray (highest height). The y-axis of (iii) and (iv) is the $\log(\text{number of participants with a heteroplasmy at that position} + 1)$. Chromosomal coordinates are displayed outside the circos plot (m.1-16569), and tracks iii-v reproduces tracks 1-3 in Hong, Battle, *et al*³. **(b)** Distribution of MLC scores for heteroplasmic variants with an allele count of 1 in UK Biobank ($n=4237$). The spike at MLC=0 includes synonymous variants. **(c)** Distribution of MLC score for variants absent in UK Biobank, at invariant sites ($n=15688$). The spike at MLC=1 includes stop gain variants. The height of the bar in (b-c) represent the variant number, plots in (b-c) display a square root transformed y-axis. UK Biobank variant data is per Hong, Battle, *et al*³.

10. Population databases

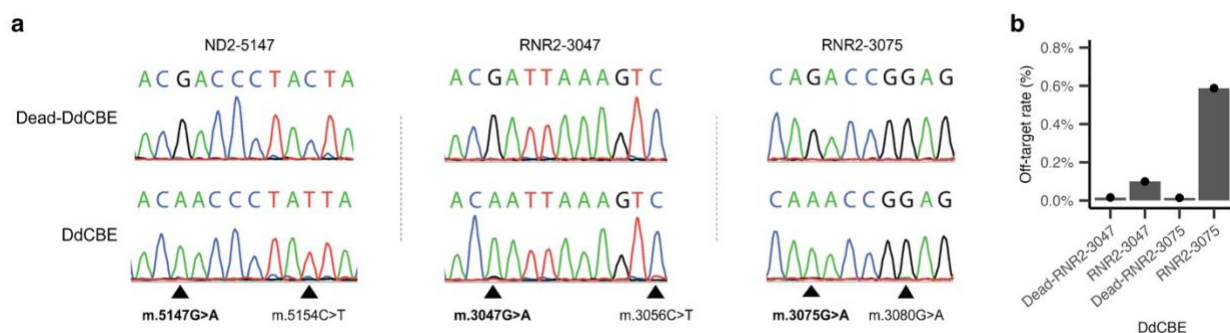
While this manuscript focuses on analysis of mitochondrial genome constraint in gnomAD, data from other publicly accessible mitochondrial population databases were also analyzed, namely HelixMTdb and MITOMAP. The population databases used in this manuscript are described in Supplementary Table 3; additional details are also provided in the main Methods. Constraint was only assessed in databases with heteroplasmy data (i.e. gnomAD with replication in HelixMTdb), while allele frequency data from gnomAD, HelixMTdb and MITOMAP was used only to assess MLC scores. Therefore, all constraint analyses presented in this manuscript use observed data only from gnomAD, with the exception of the replication analyses using HelixMTdb data.

Supplementary Table 3: Overview of the mitochondrial population databases analyzed. The sequence data type, tissue source, number of sequences, population data used, and their application is listed. Note that heteroplasmy data is not available for MITOMAP.

Database	Sequence data type, tissue source, sequence number	Population data used	Data application
gnomAD ¹⁰	- Genome sequencing - Mostly blood $n=56,434$	Maximum heteroplasmy	Constraint: gnomAD maximum heteroplasmy used for all constraint analyses, except ED Fig. 2e
		Homoplasmic allele frequency	Assessment of MLC scores: allele frequency used to compare to MLC (ED Figs. 8b & 9e, h)
HelixMTdb ²⁹	- Proprietary ‘exome plus’ sequencing - Saliva $n=195,983$	Maximum heteroplasmy	Constraint: HelixMTdb maximum heteroplasmy only used for the replication analysis (ED Fig. 2)
		Homoplasmic allele frequency	Assessment of MLC scores: allele frequency used to compare to MLC (ED Figs. 8c & 9e, i)
MITOMAP ³⁰	- GenBank sequences - Mixed tissue source $n=56,910$	Allele frequency	Assessment of MLC scores: allele frequency used to compare to MLC (ED Figs. 8d & 9e, j)

11.2 On-target and bystander base editing

On-target base editing was quantified from Sanger sequencing, as follows. DNA was extracted from cells transfected with DdCBE plasmids using PureLink Genomic DNA Mini Kit (Invitrogen) per manufacturer's protocol. Targeted PCR with Q5 polymerase (NEB) was done using primers in Supplementary Table 7, with an initial step of 3 min at 98 °C, followed by 35 cycles of 8 s at 98 °C, 15 s at 66 or 67 °C and 15 s at 72 °C, with a final 5 min at 72 °C. Sanger sequencing of PCR products was done by Quintarabio or Genewiz (US) and on-target and bystander editing was quantified using EditR³³. On-target editing >95% was achieved in HEK293T cells for each target, with range 56-98% and mean 87% across targets and replicates. RNR2-3047 and RNR2-3075 had a single bystander in the editing window with low editing (5-10%), and ND2-5147 had a second synonymous edit at similar levels to the on-target (Supplementary Fig. 10a). PCR products from cells with >95% editing per EditR were subject to next generation sequencing by Genewiz. Primer sequences were trimmed from the FASTQ using Cutadapt³⁴, and reads were aligned to hg38 reference mtDNA using BWA-MEM³⁵. Coverage of 1000-5000x was obtained. Pysam pileup (<https://github.com/pysam-developers/pysam>) was used to produce a count table with base frequency at the target sites, with a base quality filter of 30 applied. This validated on-target editing of >95%, in line with the EditR quantification of the Sanger sequences (Supplementary Table 6).



Supplementary Figure 10: Bystander edits and off-target rates. (a) Sanger sequencing chromatograms showing on-target (in bold) and bystander edits. ND2-5147 results in a synonymous mutation, while RNR2-3047 and RNR2-3075 produce target mutations in highly constrained sites. Note the bystander in ND2-5147 is another synonymous mutation. **(b)** Off-target editing rates in HEK293T cells transfected with DdCBE base editor or dead DdCBE control. The height of the bar represents a measurement in $n=1$ biological replicate.

Supplementary Table 6: Next generation sequencing validation of base editing. C>T edits were quantified in PCR amplicons from a base editing experiment by both EditR analysis of Sanger sequencing and next generation sequencing analysis, showing highly similar results.

Target	Base editing from Sanger sequencing	Base editing from next generation sequencing	Coverage depth of next generation sequencing
m.3047	97.8%	97.2%	4698x
m.3075	96.0%	97.9%	2720x
m.5147*	100%	98.6%	1089x

*The samples used for this editing quantification were not used in the functional experiments.

11.3 Off-target editing

Off-target base editing was quantified from mtDNA sequencing, as follows. DNA was extracted from cells transfected with DdCBE plasmids using PureLink Genomic DNA Mini Kit (Invitrogen) per manufacturer's protocol. Long-range PCR using Phusion Master Mix with GC Buffer (Thermo Scientific) was used to amplify the mtDNA as a single amplicon; primers and barcodes used are listed in Supplementary Table 7. The following cycling steps were used: an initial step of 30 s at 98 °C, followed by 10 cycles of 30 s at 98 °C, 30 s at 63 °C and 9 min at 72 °C and then 10 cycles of 30 s at 98 °C, 30 s at 67 °C and 9 min at 72 °C, with a final step of 10 min at 72 °C. Purification, PacBio library preparation and HiFi sequencing was done by the Yale Center for Genome Analysis using a PacBio Sequel II system. FASTQ were demultiplexed using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/download.html), and reads aligned to hg38 reference mtDNA using minimap2³⁶. Mean coverage of 30,000-40,000x was obtained. Pysam pileup (<https://github.com/pysam-developers/pysam>) was used to produce a count table with base frequency at each position; bases with quality score <30, <100x coverage and within the editing window were excluded. C:G-to-T:A variants in both edited and dead controls at heteroplasmy >10%, and variants at <0.1% heteroplasmy, were also excluded. The mean off-target rate was calculated as described³¹ (i.e. total reads with C:G-to-T:A mutations / total reads for all C•G). RNR2-3047 had a low off-target rate comparable to previous reports (0.10%)^{32,37}, while RNR2-

3075 had a higher rate (0.59%) which is likely related to its high on-target rate³⁷ (Supplementary Fig. 10b). The off-target rate in the dead controls (~0.015%) was consistent with expectation^{32,37}.

Supplementary Table 7: Primers used for sequencing. Primers for m.3047/m.3075 and m.5147 targets were used for Sanger sequencing, while primers for the entire mtDNA was used for long-read sequencing for off-target assessment; the barcodes used for the latter are in the footnote.

Target	Forward primer	Reverse primer
m.3047 / m.3075	5' TCACCAGTCAAAGCGAACTACT	5' GGGTACAATGAGGAGTAGGAGGTT
m.5147	5' CCAAATCTCTCCCTCACTAAACG	5' AGGTAGGAGTAGCGTGGTAAG
Entire mtDNA ¹	5' AATAAGACATCACGATGGATCACA	5' TAAGGGGAACGTGTGGGCTATTT

¹The following barcodes were added to the 5' end of the F or R primer for multiplex sequencing: F1-barcode 5'CAGATATC for RNR2-3047 and RNR2-3075, F2-barcode 5'GTGCACGT for Dead-RNR2-3047 and Dead-RNR2-3075, R1-barcode 5'GATTCCAC for RNR2-3047 and Dead-RNR2-3047, and R2-barcode 5'TACGATTC for RNR2-3075 and Dead-RNR2-3075.

Supplementary Discussion

1. Extended discussion on simulation of germline mtDNA mutation and heteroplasmy

We adapted a computational framework developed by Colnaghi et al²⁶ to simulate germline mtDNA mutation and heteroplasmy drift across generations, and used it to validate a correlation between mutation rates and maximum heteroplasmy for neutral mutations. The validity of our adaptation was supported by reproducing the results of an analysis reported by Colnaghi et al²⁶, which assessed the heteroplasmy distribution for five different bottleneck sizes in offspring who inherited a mutation with heteroplasmy 0.1 (Supplementary Fig. 4a). Changes in heteroplasmy levels between generations was also similar to real-world data reported by Wei et al from ~1500 mother-offspring pairs¹³, further supporting the utility of our model (Supplementary Fig. 4b). We then evaluated maximum heteroplasmy across five generations for 10 different mutation rates, using a starting heteroplasmy of 0 for generation 0 (Supplementary Fig. 4c). We used bootstrapping to assess the maximum heteroplasmy distribution at generation five for each mutation rate tested. This showed an approximately linear correlation between the mutation rate and the median maximum heteroplasmy for neutral mutations, such that higher mutation rates had higher maximum heteroplasmy, supporting our application of the mutational model for the estimation of maximum heteroplasmy values (Supplementary Fig. 4d). This is consistent with conclusions drawn by Ju et al, who found that the expected number of neutral mitochondrial mutations drifting to homoplasmy increases linearly with mutation rate in tumor cells⁵. Maximum heteroplasmy is thus influenced by mutation rate, as demonstrated in our simulation; it is also influenced by selection which reduces the heteroplasmy of deleterious variants below a ‘disease threshold’. Under this premise, we can use mutation rates to calculate expected maximum heteroplasmy and infer that when the observed is less than expected (i.e. a low observed:expected ratio) that this reflects negative selection (i.e. constraint).

2. Extended discussion on missense constraint in the mitochondrial vs nuclear genome

The mean gene missense OEUF in the mtDNA is lower than that for nuclear genes (mean nuclear gene missense OEUF 0.98 with mean observed:expected ratio 0.85). While these OEUF values are calculated differently⁸, most mtDNA-encoded proteins show similar constraint to nuclear genes in the lowest decile of nuclear missense OEUF values (with OEUF range 0.05 - 0.68 for

canonical transcripts), which includes many genes that are essential in cells or required for development or viability in mice³⁸ (22% in lowest decile vs 12% in all genes). Nuclear protein genes with similar missense constraint to most mtDNA protein genes are also enriched in genes associated with human developmental disorders caused by dominant missense mutations, with 51% of these disease genes falling in the lowest missense OEUF decile despite representing only 1% of all genes³⁹. This supports that the marked depletion of expected missense variants in most mtDNA proteins reflects their essential role in germ cell and embryonic development, and the high efficiency of mechanisms for selecting against deleterious mtDNA variants during development⁴⁰.

3. Extended discussion on regional missense constraint within protein genes

Manual inspection of regional missense constraint within protein structures revealed multiple examples of their topological clustering in 3D space. This includes for complex I subunits MT-ND1 (Fig. 2b) and MT-ND6 (Extended Data Fig. 5e), and complex III subunit MT-CYB (Extended Data Fig. 5d). Review of the literature further revealed that several of these topological clusters of regional constraint were notable for encoding residues of functional significance, as were other areas of regional constraint. For example MT-ND6 plays a key role in the transition between the ‘open’ and ‘closed’ states of complex I, which is critical for its catalytic activity⁴¹. Notably, the areas of regional constraint within this protein encode or are located in close proximity to residues critical for this function including within the π -bulge and interface-forming loop⁴¹ (Extended Data Fig. 5e). Complex III is responsible for electron transfer from ubiquinol to cytochrome c as part of oxidative phosphorylation, and it is MT-CYB that catalyzes this process. The four areas of regional constraint in MT-CYB cluster to the sites of the heme molecules and nearby quinone oxidation and reduction sites, which are critical for this function^{42,43} (Extended Data Fig. 5d). Residues in the ef loop (p.252-268), PEWY sequence (p.269-283) and cd1 helix (p.136-152) which are also critical for electron transfer⁴³ also overlap regional constraint (Supplementary Dataset 2). The one area of regional constraint identified in complex V subunit MT-ATP6 is also notable for being located at the interface with the c-ring, within an area critical for proton translation and ATP synthesis⁴⁴. Collectively, these observations support that regional missense constraint identifies areas critical for protein function.

4. Extended discussion on regional constraint within rRNA genes

There are only a handful of confirmed pathogenic variants in the mitochondrial rRNA genes, and therefore the strong negative selection observed against rRNA variation is particularly striking. However, this can be reconciled with the fact that the rRNAs have an essential role in the mitochondrial ribosome (mitoribosome) and translation. In support of this, the most regionally constrained interval in a rRNA gene (m.3028-3081 in MT-RNR2) encodes the A-loop which interacts with the 3' end of an aminoacylated tRNA in the A-site⁴⁵. This the first binding site for a tRNA in the mitoribosome during protein synthesis, and therefore likely one of the most functionally important areas of the rRNA. This area of regional constraint has an OEUF value of 0.07 (Supplementary Dataset 2), which is similar to pLoF OEUF gene values, supporting that many variants in this region are deleterious. Indeed, no homoplasmic variants were observed in this region in the gnomAD and HelixMTdb population databases collectively comprising data from up to ~250,000 individuals^{10,29}. Only five variants in this region were reported in MITOMAP³⁰; these were each reported in only one sample, and manual inspection showed that they were enriched in sample types with increased tolerance to deleterious variation and or lower quality (m.3033T>C in sample with GenBank ID LS998724.1 deposited by a Department of Pediatrics Genetic Diagnostic Laboratory, m.3058T>C in KJ735669.1 from cells, and m.3077C>T in MH043584.1 from ancient DNA)³⁰. To contextualize this observation, a similarly sized region in MT-RNR2 (m.2220-2270) has an OEUF value of 0.80 and >40 homoplasmic variants reported in gnomAD and HelixMTdb databases^{10,29}; MITOMAP also has >40 variants in this region³⁰. This example showcases a functionally critical rRNA region that may harbor overlooked deleterious variation, and supports our evidence of constraint within rRNAs, particularly within MT-RNR2 which has not had any *bona fide* pathogenic variants reported.

5. Extended discussion on local constraint score across the non-coding control region

The signal of constraint across the artificial chromosome break at m.16569-1 is unexpected (Extended Data Fig. 8a). Pipelines which use the linearized mtDNA reference sequence can be vulnerable to undercalling variants around the break; however the Mutect2 pipeline utilized for gnomAD used a shifted alignment to specifically avoid this issue¹⁰. The artificial chromosome break also lies within the OriB-OriH region, which is known to have a different mutational signature^{5,19} and was handled separately by our model (see Supplementary Methods). Future work

is required to clarify if mutational properties and or variant calling challenges unique to this region could be contributing to this signal of constraint. It is notable however that all of the other signals of constraint within the non-coding control region overlap annotated or unannotated regions previously implicated as functionally important (Extended Data Fig. 8a), including a region of unknown function previously reported to be depleted of heteroplasmies¹³. Two of these regions were only recently established to have an essential role in mitochondrial transcription, including the conserved sequence block 3⁴⁶ and the second light strand promoter⁴⁷, supporting that studies into the functional role of the region around the artificial chromosome break may also be warranted. Indeed, this region lies downstream of the origin of heavy-strand replication and nearby to the 5' ends of 7S DNA, a linear strand of DNA which forms a triple-stranded structure with the D-loop which is hypothesized to have functional significance⁴⁸.

6. Extended discussion on the local constraint scores of disease-associated variation

The mtDNA local constraint (MLC) scores of disease-associated variation shows enrichment of pathogenic variants in the higher score quartiles (Fig. 4c). However, some pathogenic variants have low scores (Extended Data Fig. 9a-d). There are several possible explanations for this. Firstly, the MLC score incorporates data from neighboring bases to measure local tolerance to variation, such that the bases in the closest proximity will contribute more of the signal. Therefore, deleterious variation could be assigned low scores if they reflect the mutation of a single critical residue or base in a neighborhood that is more tolerant of variation. This also means that positions with the highest scores are more likely to lie in sites where multiple neighboring residues/bases are critical for function. Accordingly, a benign variant could have a high score due to being in a neighborhood which is largely intolerant of base/amino acid substitution. Secondly, for missense and RNA/non-coding variants the same score is assigned for different variants at the same position, which therefore doesn't take the type of amino acid or base change into account. It's possible that a specific substitution has deleterious effects where others in the local area are tolerated. Thirdly, the heteroplasmy level that the deleterious variation is tolerated at could affect the measure of selection, such that a neighborhood that can tolerate deleterious variation at homoplasmy will have a lower score than those where strong selection occurs against lower heteroplasmy variants (e.g. where homoplasmy may not be compatible with life). This is supported by the observation that confirmed disease-associated variants reported to cause disease at homoplasmy have an increased

density of scores <0.75 compared to those with a disease association at heteroplasmy only, per their status in MITOMAP (Extended Data Fig. 9c); a trend which was strengthened following manual curation of heteroplasmy and homoplasmy status the literature (Methods) (Extended Data Fig. 9d). The MLC score is therefore positioned to be particularly useful for capturing the effect of deleterious heteroplasmic variation.

7. Extended discussion on the association between the MSS and blood cell counts

Previous studies demonstrated an association between mtDNA copy number and blood cell counts⁴⁹, and we likewise observed an association between heteroplasmy and blood cell counts, specifically for neutrophils and platelets, using a linear regression model adjusted for age, sex, and smoking status (Extended Data Table 1). Application of the mtDNA local constraint score sum (MSS) metric supported that deleterious mtDNA variation likely plays a causal role in platelet counts, and a non-causal role in neutrophil counts since neutrophil count was only significantly associated with heteroplasmy count while platelet count was only significantly associated with MSS (Extended Data Table 1). This is in line with mtDNA having a critical role in platelets due to their lack of nuclear DNA⁵⁰, and with the diminished role of mitochondria in energy generation and other cellular functions in neutrophils⁵¹. The results showed a positive correlation between heteroplasmy count and neutrophil count, as well as between the MSS and platelet count (Extended Data Table 1). The latter data point suggests that an increased burden of deleterious mtDNA variation leads to increased platelet count. While initially counterintuitive, this may reflect a possible compensatory mechanism. Support for this in the literature includes data showing that an increase in mitochondrial reactive oxygen species (a marker and consequence of mitochondrial dysfunction) can lead to platelet biogenesis⁵², and evidence that platelet apoptosis can paradoxically lead to a higher proportion of younger platelets⁵³. Further work is required to elucidate the possible mechanisms underlying these data.

8. Extended discussion on constraint in genes and regions overlooked in disease analyses

Strikingly, we identified constraint in genes and regions commonly overlooked in disease analyses. This includes the rRNA genes, which are critical for translation but often not evaluated due to a lack of *in silico* tools, as well as non-coding elements involved in mtDNA replication or transcription where variation is typically assumed benign. Although surprising, these data are

supported by studies reporting reduced transmission or frequency of variants in these loci^{10,13,18,40,54}, indicating they likely harbor pathogenic variation. We also identified strong depletion of expected variation in protein genes with few pathogenic variants reported, such as *MT-CO1* which harbored some of the most missense constrained sites in the mtDNA and yet has no ‘confirmed’ pathogenic missense³⁰. This could reflect a bias towards the study of genes that already have well-established pathogenic variants, since genes with the most ‘confirmed’ pathogenic variants are enriched with those most frequent in the population^{10,30}, and or an increased burden of deleterious variants not compatible with life. Our functional data showing the severe impact of variants in highly constrained sites in *MT-RNR2*, a rRNA gene typically overlooked in disease analyses, support that these overlooked sites harbor variation that is relevant for health and disease.

9. Extended discussion on selection against variants in the mtDNA

Selection against mtDNA variants is complex, representing a combination of forces at an organismal (i.e. reduced survival and reproductive fitness), cellular (i.e. germline or somatic), and even organelle level (i.e. mitochondrial)¹. While additional research is needed to tease apart the contribution of each to constraint in the human population, the near absence of expected loss of function variation supports that mtDNA variants with the most severe impact are not compatible with life and thus removed by selection in the germline or *in utero*. This demonstrates the high efficiency of mechanisms for preventing accumulation of deleterious mtDNA variants in the population, necessary given its high mutation rate^{17,40}. Accordingly, most pathogenic variants characterized to date will have only mild to moderate effects, given they are compatible with life but not health⁵⁵; a point which has implications for *in silico* predictors trained on reported pathogenic variants. Studies that clarify the impact of variants on germ cell and *in utero* development will thus be needed for the characterization of all deleterious mtDNA variants.

Supplementary References

- 1 Stewart, J. B. & Chinnery, P. F. Extreme heterogeneity of human mitochondrial DNA from organelles to populations. *Nature reviews. Genetics* **22**, 106-118, doi:10.1038/s41576-020-00284-x (2021).
- 2 McCormick, E. M. *et al.* Specifications of the ACMG/AMP standards and guidelines for mitochondrial DNA variant interpretation. *Hum. Mutat.* **41**, 2028-2057, doi:10.1002/humu.24107 (2020).
- 3 Hong, Y. S. *et al.* Deleterious heteroplasmic mitochondrial mutations are associated with an increased risk of overall and cancer-specific mortality. *Nat Commun* **14**, 6113, doi:10.1038/s41467-023-41785-7 (2023).
- 4 Michaelson, J. J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**, 1431-1442, doi:10.1016/j.cell.2012.11.019 (2012).
- 5 Ju, Y. S. *et al.* Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *eLife* **3**, doi:10.7554/eLife.02935 (2014).
- 6 Zheng, W., Khrapko, K., Coller, H. A., Thilly, W. G. & Copeland, W. C. Origins of human mitochondrial point mutations as DNA polymerase gamma-mediated errors. *Mutat. Res.* **599**, 11-20, doi:10.1016/j.mrfmmm.2005.12.012 (2006).
- 7 Gustafson, M. A., Sullivan, E. D. & Copeland, W. C. Consequences of compromised mitochondrial genome integrity. *DNA repair* **93**, 102916, doi:10.1016/j.dnarep.2020.102916 (2020).
- 8 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443, doi:10.1038/s41586-020-2308-7 (2020).
- 9 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).
- 10 Laricchia, K. M. *et al.* Mitochondrial DNA variation across 56,434 individuals in gnomAD. *Genome Res.* **32**, 569-582, doi:10.1101/gr.276013.121 (2022).
- 11 Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. *Nat. Genet.* **52**, 208-218, doi:10.1038/s41588-019-0572-y (2020).
- 12 Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457-465, doi:10.1038/290457a0 (1981).

- 13 Wei, W. *et al.* Germline selection shapes human mitochondrial DNA diversity. *Science* **364**, eaau6520, doi:10.1126/science.aau6520 (2019).
- 14 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, doi:10.1371/journal.pgen.1003794 (2013).
- 15 Ludwig, L. S. *et al.* Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell* **176**, 1325-1339.e1322, doi:10.1016/j.cell.2019.01.022 (2019).
- 16 Rebolledo-Jaramillo, B. *et al.* Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 15474-15479, doi:10.1073/pnas.1409328111 (2014).
- 17 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, doi:10.1073/pnas.1906331116 (2019).
- 18 Li, M. *et al.* Transmission of human mtDNA heteroplasmy in the Genome of the Netherlands families: support for a variable-size bottleneck. *Genome Res.* **26**, 417-426, doi:10.1101/gr.203216.115 (2016).
- 19 Yuan, Y. *et al.* Comprehensive molecular characterization of mitochondrial genomes in human cancers. *Nat. Genet.* **52**, 342-352, doi:10.1038/s41588-019-0557-x (2020).
- 20 pfeliciano@simonsfoundation.org, S. C. E. a. & Consortium, S. SPARK: A US cohort of 50,000 families to accelerate autism research. *Neuron* **97**, 488-493, doi:10.1016/j.neuron.2018.01.015 (2018).
- 21 Wang, Y. *et al.* Association of mitochondrial DNA content, heteroplasmies and inter-generational transmission with autism. *Nat Commun* **13**, 3790, doi:10.1038/s41467-022-30805-7 (2022).
- 22 Van Oven, M. PhyloTree Build 17: Growing the human mitochondrial DNA tree. *Forensic Science International: Genetics Supplement Series* **5**, e392-e394 (2015).
- 23 Diedenhofen, B. & Musch, J. cocor: a comprehensive solution for the statistical comparison of correlations. *PLoS One* **10**, e0121945, doi:10.1371/journal.pone.0121945 (2015).

- 24 Stocco, A. & Coppedè, F. Mitochondrial DNA methylation and human diseases. *Int. J. Mol. Sci.* **22**, doi:10.3390/ijms22094594 (2021).
- 25 Bicci, I., Calabrese, C., Golder, Z. J., Gomez-Duran, A. & Chinnery, P. F. Single-molecule mitochondrial DNA sequencing shows no evidence of CpG methylation in human cells and tissues. *Nucleic Acids Res.* **49**, 12757-12768, doi:10.1093/nar/gkab1179 (2021).
- 26 Colnaghi, M., Pomiankowski, A. & Lane, N. The need for high-quality oocyte mitochondria at extreme ploidy dictates mammalian germline development. *eLife* **10**, doi:10.7554/eLife.69344 (2021).
- 27 Samocha, K. E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *BioRxiv*, doi:10.1101/148353 (2017).
- 28 Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS computational biology* **6**, e1001025, doi:10.1371/journal.pcbi.1001025 (2010).
- 29 Bolze, A. *et al.* A catalog of homoplasmic and heteroplasmic mitochondrial DNA variants in humans. *BioRxiv*, doi:10.1101/798264 (2020).
- 30 Lott, M. T. *et al.* mtDNA Variation and Analysis Using MITOMAP and MITOMASTER. *Current protocols in bioinformatics* **1**, 1.23.21-21.23.26, doi:10.1002/0471250953.bi0123s44 (2013).
- 31 Mok, B. Y. *et al.* A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. *Nature* **583**, 631-637, doi:10.1038/s41586-020-2477-4 (2020).
- 32 Mok, B. Y. *et al.* CRISPR-free base editors with enhanced activity and expanded targeting scope in mitochondrial and nuclear DNA. *Nat. Biotechnol.* **40**, 1378-1387, doi:10.1038/s41587-022-01256-8 (2022).
- 33 Kluesner, M. G. *et al.* EditR: A Method to Quantify Base Editing from Sanger Sequencing. *CRISPR J* **1**, 239-250, doi:10.1089/crispr.2018.0014 (2018).
- 34 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10-12 (2011).
- 35 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).

- 36 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).
- 37 Silva-Pinheiro, P. *et al.* A library of base editors for the precise ablation of all protein-coding genes in the mouse mitochondrial genome. *Nat Biomed Eng* **7**, 692-703, doi:10.1038/s41551-022-00968-1 (2023).
- 38 Cacheiro, P. *et al.* Human and mouse essentiality screens as a resource for disease gene discovery. *Nat Commun* **11**, 655, doi:10.1038/s41467-020-14284-2 (2020).
- 39 Firth, H. V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* **84**, 524-533, doi:10.1016/j.ajhg.2009.03.010 (2009).
- 40 Floros, V. I. *et al.* Segregation of mitochondrial DNA heteroplasmy through a developmental genetic bottleneck in human embryos. *Nat. Cell Biol.* **20**, 144-151, doi:10.1038/s41556-017-0017-8 (2018).
- 41 Kampjut, D. & Sazanov, L. A. The coupling mechanism of mammalian respiratory complex I. *Science* **370**, abc4209, doi:10.1126/science.abc4209 (2020).
- 42 Xia, D. *et al.* Structural analysis of cytochrome bc₁ complexes: implications to the mechanism of function. *Biochim. Biophys. Acta* **1827**, 1278-1294, doi:10.1016/j.bbabo.2012.11.008 (2013).
- 43 Rajagukguk, S. *et al.* Effect of mutations in the cytochrome b ef loop on the electron-transfer reactions of the Rieske iron-sulfur protein in the cytochrome bc₁ complex. *Biochemistry* **46**, 1791-1798, doi:10.1021/bi062094g (2007).
- 44 Su, X. *et al.* The pathogenic m.8993 T > G mutation in mitochondrial ATP6 gene prevents proton release from the subunit c-ring rotor of ATP synthase. *Hum. Mol. Genet.* **30**, 381-392, doi:10.1093/hmg/ddab043 (2021).
- 45 Koripella, R. K., Sharma, M. R., Risteff, P., Keshavan, P. & Agrawal, R. K. Structural insights into unique features of the human mitochondrial ribosome recycling. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 8283-8288, doi:10.1073/pnas.1815675116 (2019).
- 46 Zhu, X. *et al.* Non-coding 7S RNA inhibits transcription via mitochondrial RNA polymerase dimerization. *Cell*, doi:10.1016/j.cell.2022.05.006 (2022).
- 47 Tan, B. G. *et al.* The human mitochondrial genome contains a second light strand promoter. *Mol. Cell* **82**, 3646-3660.e3649, doi:10.1016/j.molcel.2022.08.011 (2022).

- 48 Nicholls, T. J. & Minczuk, M. In D-loop: 40 years of mitochondrial 7S DNA. *Exp. Gerontol.* **56**, 175-181, doi:10.1016/j.exger.2014.03.027 (2014).
- 49 Liu, X. *et al.* Association of mitochondrial DNA copy number with cardiometabolic diseases. *Cell Genomics* **1**, doi:10.1016/j.xgen.2021.100006 (2021).
- 50 Melchinger, H., Jain, K., Tyagi, T. & Hwa, J. Role of Platelet Mitochondria: Life in a Nucleus-Free Zone. *Frontiers in cardiovascular medicine* **6**, 153, doi:10.3389/fcvm.2019.00153 (2019).
- 51 Maiani, N. A. *et al.* Functional characterization of mitochondria in neutrophils: a role restricted to apoptosis. *Cell Death and Differentiation* **11**, 143-153, doi:10.1038/sj.cdd.4401320 (2004).
- 52 Poirault-Chassac, S. *et al.* Mitochondrial dynamics and reactive oxygen species initiate thrombopoiesis from mature megakaryocytes. *Blood Adv* **5**, 1706-1718, doi:10.1182/bloodadvances.2020002847 (2021).
- 53 Tyagi, T. *et al.* A guide to molecular and functional investigations of platelets to bridge basic and clinical sciences. *Nat Cardiovasc Res*, doi:10.1038/s44161-022-00021-z (2022).
- 54 Voets, A. M. *et al.* Large scale mtDNA sequencing reveals sequence and functional conservation as major determinants of homoplasmic mtDNA variant distribution. *Mitochondrion* **11**, 964-972, doi:10.1016/j.mito.2011.09.003 (2011).
- 55 Fan, W. *et al.* A mouse model of mitochondrial disease reveals germline selection against severe mtDNA mutations. *Science* **319**, 958-962, doi:10.1126/science.1147786 (2008).

Supplementary Datasets

Supplementary Dataset 1: Constraint metrics for human mtDNA genes. Gene symbol and coordinates, and the observed and expected sum maximum heteroplasmy of SNVs in the gene in gnomAD v3.1, observed:expected ratio, and lower and upper bounds of the 90% confidence interval (CI) around the ratio are listed, including for each major class of protein variation (synonymous, missense, stop gain) for protein genes.

Supplementary Dataset 2: Intervals of regional constraint in mitochondrial proteins and rRNA genes. Gene symbol and coordinates of regional constraint, and the observed and expected sum maximum heteroplasmy of SNVs in the region in gnomAD v3.1, observed:expected ratio, and lower and upper bounds of the 90% confidence interval (CI) around the ratio are listed. Note this represents regional missense constraint for protein genes, and regional constraint for RNA variants in rRNA genes.

Supplementary Dataset 3: Curated missense variants identified in cases suspected to have a primary mitochondrial disorder. Curation and classification were performed by the Victorian Clinical Genetics Service, and variant position, reference and alternate alleles, and their classification group are listed. These data are shown in Fig. 2d.

Supplementary Dataset 4: Constraint metrics for each position within the tRNA secondary structure. The tRNA secondary structure position number, and the observed and expected sum maximum heteroplasmy of SNVs at the position in gnomAD v3.1, observed:expected ratio, and lower and upper bounds of the 90% confidence interval (CI) around the ratio are listed.

Supplementary Dataset 5: Constraint metrics for annotated non-coding elements. Locus name, description and coordinates, and the observed and expected sum maximum heteroplasmy of SNVs in the element in gnomAD v3.1, observed:expected ratio, and lower and upper bounds of the 90% confidence interval (CI) around the ratio are listed.

Supplementary Dataset 6: mtDNA local constraint (MLC) base scores for every base position. The mtDNA position and its position score is listed. Note this score measures missense tolerance specifically for positions in protein genes (Methods).

Supplementary Dataset 7: mtDNA local constraint (MLC) variant scores for every single nucleotide variant. The variant position, reference and alternate alleles, consequence and MLC score are listed. Note that non-missense variants in proteins were assigned scores separately (Methods).

Supplementary Dataset 8: Neutral variants used to validate the mutational model. The variant and criteria for inclusion is listed.

Supplementary Dataset 9: Confirmed pathogenic variants with a curated disease plasmy status. Disease-associated variants with a ‘confirmed’ status and plasmy status of ‘-/+’ in MITOMAP had their plasmy status in the literature manually reviewed. The variant, its gene symbol and MITOMAP plasmy status, status after curation, and PMIDs of publications used for the curated status are listed. Note only missense and RNA variants are included. These data are shown in Extended Data Fig. 9d.

Supplementary Videos

Supplementary Video: Local constraint across the 16S rRNA encoded by MT-RNR2. The video shows the mitoribosome, a complex of proteins (blue) and the small 12S (pink) and large 16S (purple) rRNAs, which serves as the site of mitochondrial translation. The mRNA (bright green) and tRNAs occupying the A/P-site (yellow) and P/E-site (orange) are also shown. The mtDNA local constraint scores are then displayed across the 16S rRNA encoded by MT-RNR2 using a red-white-blue gradient. Dark red indicates highly constrained sites with scores close to 1, white scores around 0.5, and dark blue scores approaching 0.