Human Genetics (2019) 138:613–624
https://doi.org/10.1007/s00439-019-02009-5

ORIGINAL INVESTIGATION

Mutation signatures in germline mitochondrial genome provide insights into human mitochondrial evolution and disease

Xiwen Gu[1] · Xinyun Kang[1] · Jiankang Liu[2]

Received: 4 December 2018 / Accepted: 2 April 2019 / Published online: 9 April 2019

Abstract

Variations in mitochondrial DNA (mtDNA) have been fundamental for understanding human evolution and are causative for a plethora of inherited mitochondrial diseases, but the mutation signatures of germline mtDNA and their value in understanding mitochondrial pathogenicity remain unknown. Here, we carried out a systematic analysis of mutation patterns in germline mtDNA based on 97,566 mtDNA variants from 45,494 full-length sequences and revealed a highly non-stochastic and replication-coupled mutation signature characterized by nucleotide-specific mutation pressure (G > T > A > C) and position-specific selection pressure, suggesting the existence of an intensive mutation–selection interplay in germline mtDNA. We provide evidence that this mutation–selection interplay has strongly shaped the mtDNA sequence during evolution, which not only manifests as an oriented alteration of amino acid compositions of mitochondrial encoded proteins, but also explains the long-lasting mystery of CpG depletion in mitochondrial genome. Finally, we demonstrated that these insights may be integrated to better understand the pathogenicity of disease-implicated mitochondrial variants.

Introduction

Mitochondria are cytoplasmic organelles central for energy metabolism, biosynthesis of macromolecules, production of reactive oxygen species (ROS), and apoptosis signaling (Vyas et al. 2016). The circular mitochondrial genome (mtDNA) consists of only 16,569 base pairs, but is present in hundreds to thousands of copies in each cell and encodes 37 genes including 2 rRNAs, 22 tRNAs, and 13 proteins essential for oxidative phosphorylation (Stewart and Chinnery 2015; Wallace and Chalkia 2013).

mtDNA genetics are characterized by maternal inheritance, high mutation rate, the lack of recombination, and the presence of a germline bottleneck (Floros et al. 2018; Rebolledo-Jaramillo et al. 2014). Deleterious mtDNA mutations are responsible for a plethora of inherited mitochondrial diseases such as Leber hereditary optic neuropathy (LHON) and mitochondrial encephalomyopathy, lactic acidosis and stroke-like episodes (MELAS), and Leigh syndrome (LS) etc. (Chinnery 2015; Lightowlers et al. 2015). These diseases are rare in separation, but are among the most common and complex of all inherited diseases in combination (Lightowlers et al. 2015). According to MITOMAP (www.mitomap.org) (Lott et al. 2013; Sonney et al. 2017), a total of 680 mtDNA variants have been implicated in mitochondrial diseases, but only around 10% of them are recognized as being pathogenic by the mitochondrial research community. Thus, there exists a great gap in our understanding the pathogenicity of disease-implicated mitochondrial variants.

Mutational signature is the characteristic pattern of mutations formed by combined mutation and repair process, which has provided numerous insights into the pathogenesis of cancer (Alexandrov et al. 2013; Helleday et al. 2014). Recently, several groups have reported the mutation

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00439-019-02009-5) contains supplementary material, which is available to authorized users.

✉ Xiwen Gu
xiwen.gu@xjtu.edu.cn

✉ Jiankang Liu
j.liu@mail.xjtu.edu.cn

[1] Key Laboratory of Shaanxi Province for Craniofacial Precision Medicine Research, Clinical Research Center of Shaanxi Province for Dental and Maxillofacial Diseases, College of Stomatology, Xi'an Jiaotong University, Xi'an 710004, China

[2] Center for Mitochondrial Biology and Medicine & Douglas C. Wallace Institute for Mitochondrial and Epigenetic Information Sciences, The Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China

signature of somatic mtDNA in cancer (Ju et al. 2014; Stewart et al. 2015; Zeng et al. 2018) and aging process (Itsara et al. 2014; Kennedy et al. 2013), but the mutation signature in germline mtDNA and its impact on evolution remain less known. It is also unknown whether deciphering the germline mtDNA signature can facilitate the understanding of pathogenicity for disease-implicated mitochondrial variants.

Two challenges have hampered the investigation of mutational signature in germline mtDNA. First, mutational events in human mtDNA are traditionally identified relative to a contemporary European sequence (the rCRS sequence), which is often misleading in classifying ancestral and derived variants (Behar et al. 2012). Second, because of the short nature of mtDNA, enormous variants in human population are clustered at identical positions (Kloss-Brandstatter et al. 2011), making it difficult to distinguish multiple de novo mutation events from single historical mutation event. To date, no systematic analysis of mutation signature in human germline mtDNA has been reported (Ju et al. 2014; Stewart et al. 2015).

Here, we performed a systematic analysis of mutation signature in germline human mtDNA by distinguishing phylogeny-based "historical" variants (Pereira et al. 2011; van Oven and Kayser 2009) and frequency-based "modern" variants. Combined analysis of historical and modern variants revealed the existence of a robust and distinct germline mtDNA signature, which not only lends novel insights into human mtDNA evolution, but also has potential clinical relevance in understanding the pathogenicity of disease-implicated mtDNA variants.

Results and discussion

The catalog of germline mtDNA variants

A total of 97,566 human mtDNA variants were retrieved from the MITOMAP database (Lott et al. 2013). The majority of them (94%, 91,718 variants) were single nucleotide substitutions, affecting 51% of the mitochondrial genome (8359 positions) with massive identical variants presented in multiple haplogroups (Supplementary Fig. 1). The existence of identical variants in different haplogroups may reflect the occurrence of multiple de novo mutation and/or the spreading of ancestral mutations along progeny, making direct analysis of mutation patterns impossible. To overcome this obstacle, we stratified the mtDNA variations into "historical" (defining variants for individual haplogroup) and "modern" variants (recent de novo mutations). For "historical" variants, we used a phylogeny-based approach and extracted 11,746 haplogroup defining variants (Supplementary Table 1) from the updated Phylotree, which is based on the reconstructed ancestral genome of Mitochondrial Eve (the RSRS sequence) (Behar et al. 2012; van Oven and Kayser 2009). For "modern" variants, we developed a frequency-based method and extracted 44,334 unique variants (Supplementary Table 2). The dichotomy of historical and modern variants also enabled us to compare and cross-validate our findings.

Distinct mutation signatures in germline mitochondrial genome

Analysis of these germline substitutions revealed a high predominance of transition mutations in both historical (96%) and modern (93%) mtDNA variants (Supplementary Fig. 2), consistent with the reported paucity of ROS-associated transversion mutations in somatic mtDNA (Kauppila and Stewart 2015; Kennedy et al. 2013; Williams et al. 2013). Analysis of the coding region mutations (577-16023) by trinucleotide context revealed a highly identical mutation signature in both historical and modern variants (Fig. 1a), and across different geographical populations (Supplementary Fig. 3), together with largely identical hotspot enrichments (Fig. 1b), demonstrating the robustness and non-stochastic nature of the mutational process in germline mtDNA. Furthermore, our analysis revealed only a moderate heavy strand (H strand) bias for C > T transitions (3.4-fold for historical and 3-fold for modern variants), lack of light strand (L strand) bias for T > C transitions, and no obvious hotspot enrichment for CpG sites (Fig. 1a), greatly contrasting with the reported extreme strand bias and CpG enrichment in somatic mtDNA variants (Ju et al. 2014; Stewart et al. 2015). Thus, we conclude that the coding region mutation signature of germline mtDNA is highly robust, but distinct from somatic mtDNA.

Generation of mitochondrial signature via combined mutation and purifying selection

To gain insight into the difference between somatic and germline mtDNA variants, we further examined the D-loop region (16106-191), which is formed by incorporation of a short DNA strand known as 7S DNA (Nicholls and Minczuk 2014). We found that the D-loop region in germline mtDNA exhibited a reversed, L strand-biased C > T transitions (Supplementary Fig. 4), consistent with the observations in somatic mtDNA (Ju et al. 2014). Because mtDNA D-loop region is noncoding and overlaps with two of the most hypervariable segments (HVS1 and HVS2) (Nicholls and Minczuk 2014), we reasoned that the D-loop may be subjected to less functional constraints than the coding region. Thus, the shared mutation signature in the D-loop, but distinct mutation signature in the coding region may reflect the presence of a uniform mutation pressure plus a

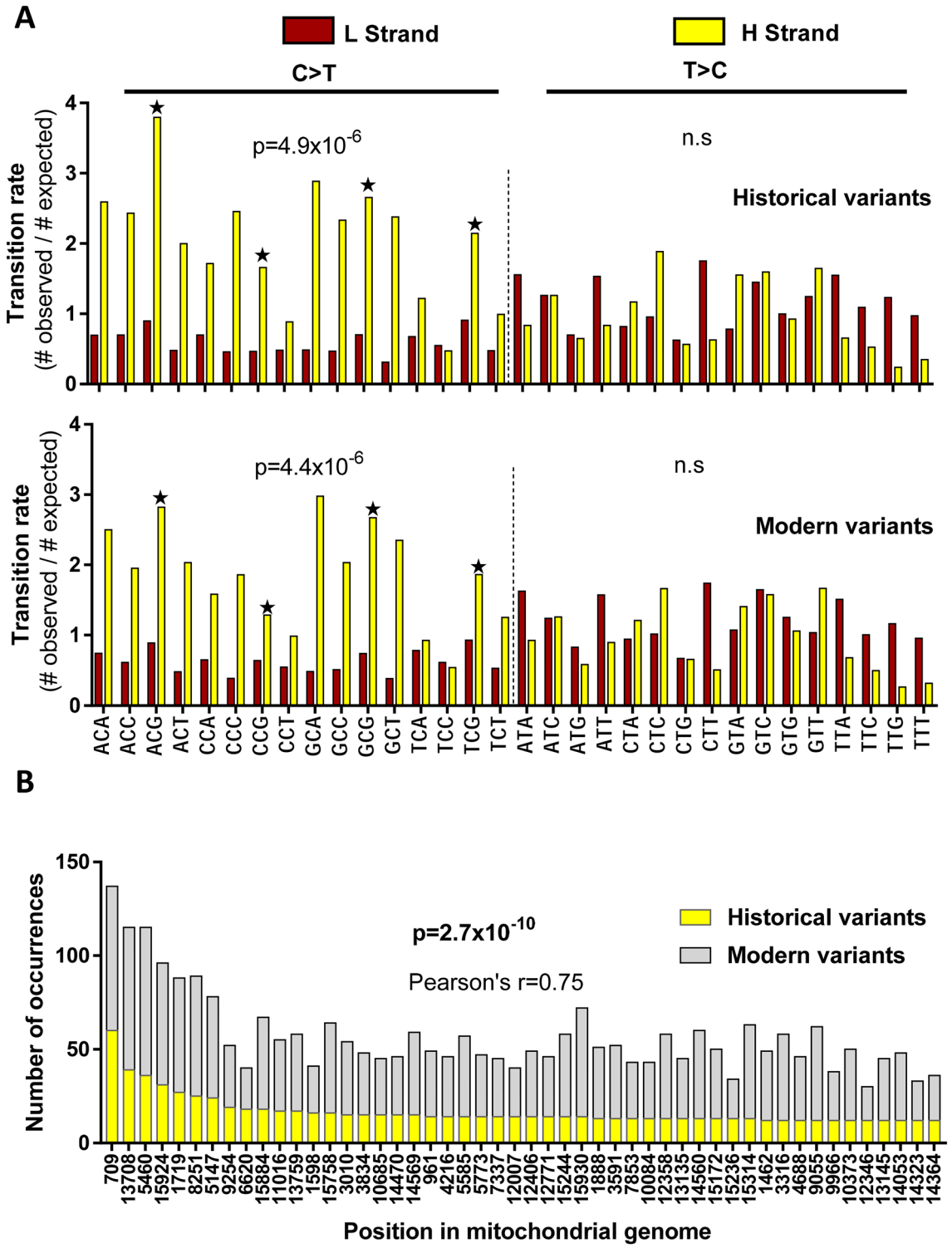


Fig. 1 Robust mutation signature in germline mtDNA. **a** Strand-specific transition signature in the coding region (577-16023) of germline mtDNA based on historical (top) and modern variants (bottom). Transition rates (# of observed/# of expected) are denoted by pyrimidine according to 64 trinucleotide contexts, with H strand and L strand patterns shown in parallel. CpG sites are highlighted by asterisk (*). C > T and T > C transitions are bordered by a dashed line with the respective significance calculated separately by Welch two-sample *t* test. *n.s.* not significant. **b** Recurrent targeting of hotspot positions in both historical and modern variants. The number of transition occurrence for the top 50 mostly targeted positions in historical variants are shown, along with their corresponding number of occurrence in modern variants

distinct selection pressure between somatic and germline mtDNA.

Transmission of germline mtDNA is distinct from somatic mtDNA by the presence of a genetic bottleneck (Floros et al. 2018; Rebolledo-Jaramillo et al. 2014), which facilitates rapid segregation and fixation of mtDNA variants via reduction of mtDNA copy number in early primordial germ cells, and the occurrence of strong purifying selection (Fan et al. 2008; Stewart et al. 2008a, b), which eliminates deleterious mtDNA variants from the female germline. To test whether the mutation signature of germline mtDNA was shaped by these processes, we divided the mtDNA coding region into six functional units (12s rRNA, 16s rRNA, tRNA, and 1st, 2nd, 3rd position of the codon triplet) and characterized their signature, respectively. As shown in Fig. 2, the third codon mutation signature of germline mtDNA exhibited an extreme strand asymmetry, with H strand-biased C > T (8.2-fold enrichment) and L strand-biased T > C transition (2.4-fold enrichment), and an absence of CpG enrichment (Fig. 2, bottom). This strength of strand asymmetry is comparable to the reported somatic mtDNA signature (Ju et al. 2014; Stewart et al. 2015). However, in all other function units, the signature of strand asymmetry in C > T was barely detectable and the signature of strand-biased T > C transitions was completely absent, accompanied by a significant reduction in the transition rate (Fig. 2). Because the transition dominated nature of mtDNA variants makes the third codon positions essentially neutral and can serve as a footprint for selection-free mutation process in germline mtDNA, the reduction of transition rate and alteration of mutation signature in protein synthesizing (rRNA and tRNA) and amino acid altering (1st and 2nd positions of codon triplet) units were hallmarks of strong purifying selection, suggesting that mutation signature in germline mtDNA is generated by the combined action of mitochondrial distinct mutation process and germline-specific purifying selection.

Replication-coupled mutation signature in germline mtDNA

Accumulated evidence suggests that the extreme mutational strand asymmetry in human mtDNA is related to the distinct mode of mtDNA replication, where the nascent H strand is synthesized first while the parental H strand is displaced and is single stranded (Ju et al. 2014; Stewart et al. 2015; Wanrooij and Falkenberg 2010). To gain further insight, we analyzed the correlation between the mutation rate and the duration of parental H strand being single stranded during replication (DssH) using neutral third codon variants in germline mtDNA. Both the rates of G > A (H strand C > T, $p = 0.003$) and T > C (H strand A > G, $p = 0.05$) transitions were significantly elevated and positively proportional to DssH (Fig. 3). This presence of H strand-specific mutation gradient relative to DssH supports that hydrolytic deamination of cytosine and adenine on the parental H strand may play an important role in mtDNA mutagenesis. In addition, the H strand-biased mutation is highly evident even at regions with the lowest DssH value (Cox1 gene, Fig. 3), suggesting that the rate and processivity of nascent H strand replication may be different from the nascent L strand synthesis, which in turn can causes lower fidelity and higher error rate during nascent H strand replication. Thus, the distinct mode of mtDNA replication may render both the parental H strand (via spontaneous deamination of cytosine and adenine) and nascent H strand (via decreased replication fidelity) vulnerable to mutagenesis, explaining the extremely H strand-biased mutation signature of mtDNA.

Further dissection of mutation and purifying selection in germline mtDNA

To better understand the relative strength of mutation and purifying selection, we stratified the mutations of 12 L strand encoded proteins into amino acid specifying codon groups and measured their respective mutation and selection pressure, with the latter inferred as the distance between the observed mutation rate and predicted selection-free mutation rate (mutation pressure). As shown in Fig. 4a, the leucine-specifying TTR codon group (1st codon position) showed almost equal level of observed and predicted transition rate, which is in agreement with our expectation because transition of TTR at the first codon position is synonymous and neutral (both TTR and CTN specify leucine). Thus, the TTR codon set served both as a negative control for selection pressure and a positive control for selection-free mutation pressure, demonstrating the feasibility of our estimation of mutation pressure. The majority of codon groups exhibited a striking reduction in the observed transition rate relative to the predicted selection-free transition rate, highlighting the strength of purifying selection. Notably, both the mutation pressure and selection pressure were exceptionally high at nucleotides GT, but were much lower at nucleotides AC (Fig. 4a, b). In addition, the selection pressure is not confined to protein-coding genes, as comparable or even higher levels of selection (ranging from 74 to 83% of mutation pressure) were also evident for tRNA and rRNA genes (Fig. 4c). Consistently, at the neutral third position in the codon triplet, the presence of multiple variants was observed for more than 90% of G/T nucleotide positions and ~ 60% of AC nucleotide positions, whereas only ~ 20% of positions were affected in the remaining protein- and RNA-coding regions in germline mtDNA (Fig. 4d). Considering human population as a whole, these data are consistent with a scenario where mitochondrial strand-biased mutation process generated an almost saturated pool of transition variants spanning the whole mtDNA, but these variants were subjected to stringent

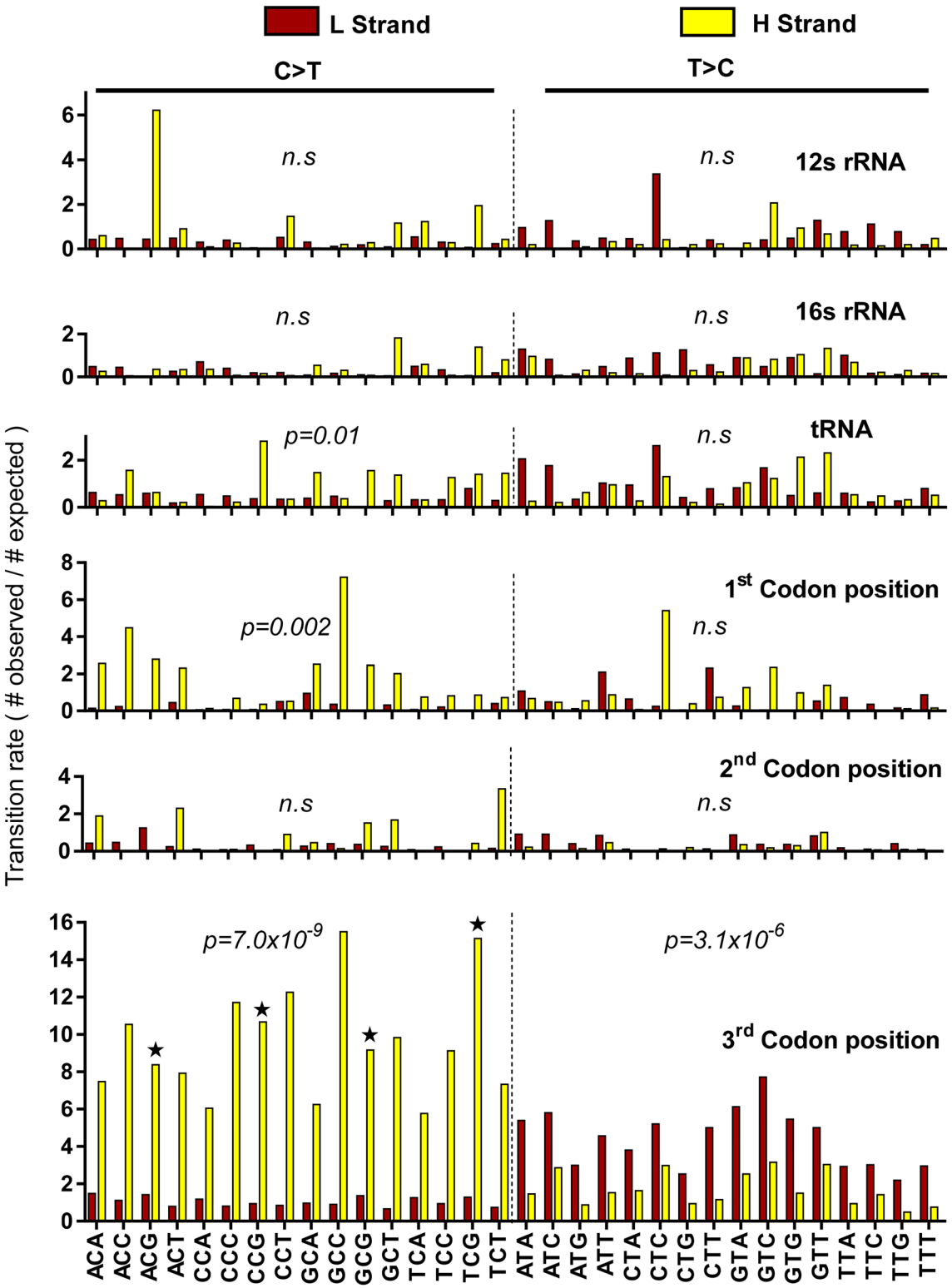


Fig. 2 Locus-specific transition signature in germline mtDNA. The pattern of transition was analyzed for 12s rRNA, 16s rRNA, tRNA, first, second, and third codon positions, respectively, using combined historical and modern variants. Transition rates are denoted by pyrimidine according to 64 trinucleotide contexts, with H strand and L strand mutation shown in parallel to reveal strand biases. C > T and T > C transitions are bordered by a dashed line with the respective significance calculated by Welch two-sample *t* test. Trinucleotide context harboring CpG sites are highlighted by single asterisk (*) at the third codon position. *n.s* not significant

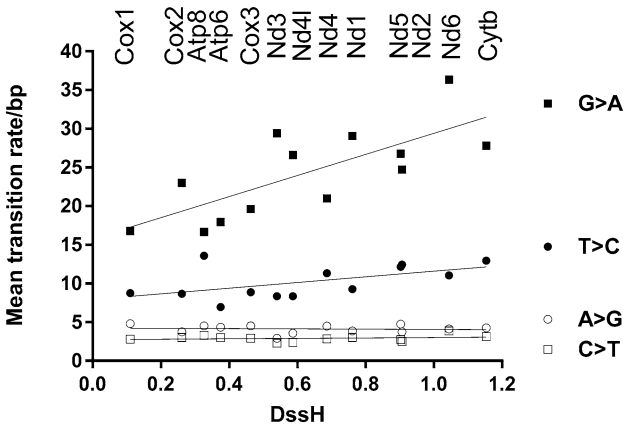


Fig. 3 Plot of gene-specific G > A, T > C, A > G, and C > T transition rates on DssH (duration of being single stranded during replication) for the neutral third codon variants in germline mtDNA. The transition rate and corresponding DssH are shown for each of the 13 mtDNA encoded protein genes. The linear regression lines are also shown. Cox1 gene has the lowest DssH value (0.11), while Cytb gene has the highest DssH value (1.15)

functional test and selection in female germline, such that the majority of deleterious variants were swept out in the population. On the other hand, as few evidence of purifying selection has been reported in somatic mtDNA variants derived from cancer and aging process (Itsara et al. 2014; Ju et al. 2014; Kennedy et al. 2013; Stewart et al. 2015; Zeng et al. 2018), the somatic and germline mtDNA variants may represent very distinct mtDNA mutation pools, with somatic variants harboring more deleterious nonsynonymous changes.

Strong shaping of mitochondrial genome by mutation and selection

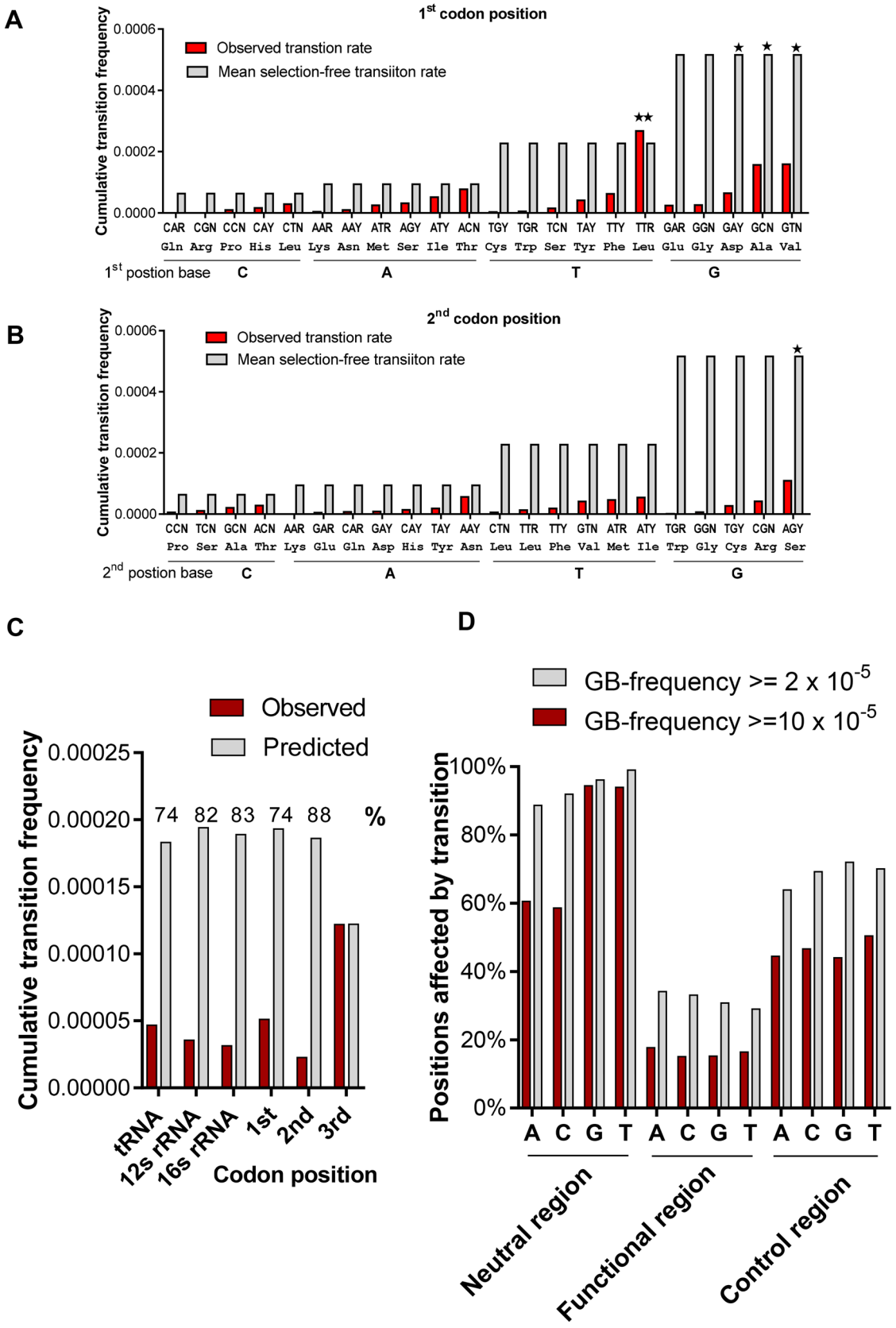
Another scenario that emerges from the germline mutation signature is the existence of both impetus and resistance for directional evolution in mitochondrial genome, with the former driven by nucleotide-specific mutation pressure and manifested as a net flow of G > A and T > C on the L strand, and the latter driven by selection pressure that counteract the directional mutation force. The directional evolution was most evident at the third codon positions where selection pressure was almost absent, leading to severe depletion of G/T nucleotides (Ju et al. 2014; Kivisild et al. 2006), which has also been intensively investigated based on the nucleotide compositional bias (Faith and Pollock 2003; Fonseca et al. 2014; Raina et al. 2005; Reyes et al. 1998). The resistance to directional evolution was evident at other functional regions where strength of purifying selection is generally high (Fig. 4). We further provided evidence that directional evolution was also operative at the functional first and second codon positions when selection pressure did not completely abrogate directional mutation, manifested as variable depletion of codons for valine, aspartic acid, alanine, and serine and the corresponding expansion of codons for methionine, isoleucine, asparagine, and threonine (Fig. 5a–d). No sign of directional evolution was observed in codons for glycine that suffer from extreme selection pressure (Fig. 5e).

Fig. 4 Mutation–selection interplay in germline mtDNA. Codon-specific analyses of mutation and selection at the first codon position **a** and second codon position, **b** for 12 mitochondrial L strand proteins are shown. Mean selection-free transition rate is estimated based on the neutral third codon position. The strength of selection pressure is visualized by the distance between the observed (red) and selection-free transition rate (gray). 22 codon families are shown along with their specifying amino acid by merging twofold and fourfold degenerate sites, with R for purine (A or G) and Y for pyrimidine (C or T) and N for any nucleotide. The selection-free TTR codon (specifying leucine) is highlighted with double asterisks (**). **c** Locus-specific mutation and selection pressure in germline mtDNA. The predicted mutation pressure was estimated based on the neutral third codon position. The number on the bar indicates the percentage of selection pressure relative to the predicted mutation pressure. **d** Positions affected by mutation in germline mtDNA for neutral region (3rd codon positions), functional region (rRNA, tRNA, 1st and 2nd codon position), and D-loop region (16024-576). Variants affecting at least one mtDNA sequence (red bar) and variants affecting five or more mtDNA sequence (gray bar) are shown in parallel

We next examined whether mutation signature in germline mtDNA could resolve the long-lasting mystery of pervasive CpG depletion in human mtDNA (Cardon et al. 1994). Both CpG and GpC sites exhibited a comparable, highly L strand-biased transition rate at neutral third positions, arguing against a role for CpG methylation (Fig. 2). CpG sites at codon position [1:2] were underrepresented (190 less) in comparison to GpC, corresponding to the low arginine usage (codon CGN) and high alanine usage (codon GCN). At codon position [2:3] and [3:1], CpG and GpC exhibited a reversed distribution, with CpG sites enriched at [3:1] but severely depleted at [2:3], and GpC sites showing the opposite (Fig. 5f). This pattern of distribution argued against a CpG-specific depletion mechanism, but can be fully explained by the mutation and selection process in germline mtDNA. Taken together, these data demonstrated that the mutation signature in germline mtDNA has strongly shaped the mtDNA during evolution.

Integrating mutation–selection interplay to understand the pathogenicity of mitochondrial variants

Lastly, we tested whether insights gained from germline mtDNA variants can be used to evaluate pathogenicity using disease-associated variants retrieved from MITOMAP. Of the 83 mtDNA variants with confirmed pathogenicity, there were 5 indels, 8 transversions, and 70 transitions (29 G > A,


A
1st codon position
Cumulative transition frequency
0.0006
0.0004
0.0002
0.0000
Observed transtion rate
Mean selection-free transiiton rate
CAR CGN CCN CAY CTN AAR AAY ATR AGY ATY ACN TGY TGR TCN TAY TTY TTR GAR GGN GAY GCN GTN
Gln Arg Pro His Leu Lys Asn Met Ser Ile Thr Cys Trp Ser Tyr Phe Leu Glu Gly Asp Ala Val
1st postion base C A T G
B
2nd codon position
Cumulative transition frequency
0.0006
0.0004
0.0002
0.0000
Observed transtion rate
Mean selection-free transiiton rate
CCN TCN GCN ACN AAR GAR CAR GAY CAY TAY AAY CTN TTR TTY GTN ATR ATY TGR GGN TGY CGN AGY
Pro Ser Ala Thr Lys Glu Gln Asp His Tyr Asn Leu Leu Phe Val Met Ile Trp Gly Cys Arg Ser
2nd postion base C A T G
C
Observed
Predicted
Cumulative transition frequency
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
74 82 83 74 88 %
tRNA
12s rRNA
16s rRNA
1st
2nd
3rd
Codon position
D
GB-frequency >= 2 x 10-5
GB-frequency >=10 x 10-5
Positions affected by transition
100%
80%
60%
40%
20%
0%
A C G T A C G T A C G T
Neutral region
Functional region
Control region

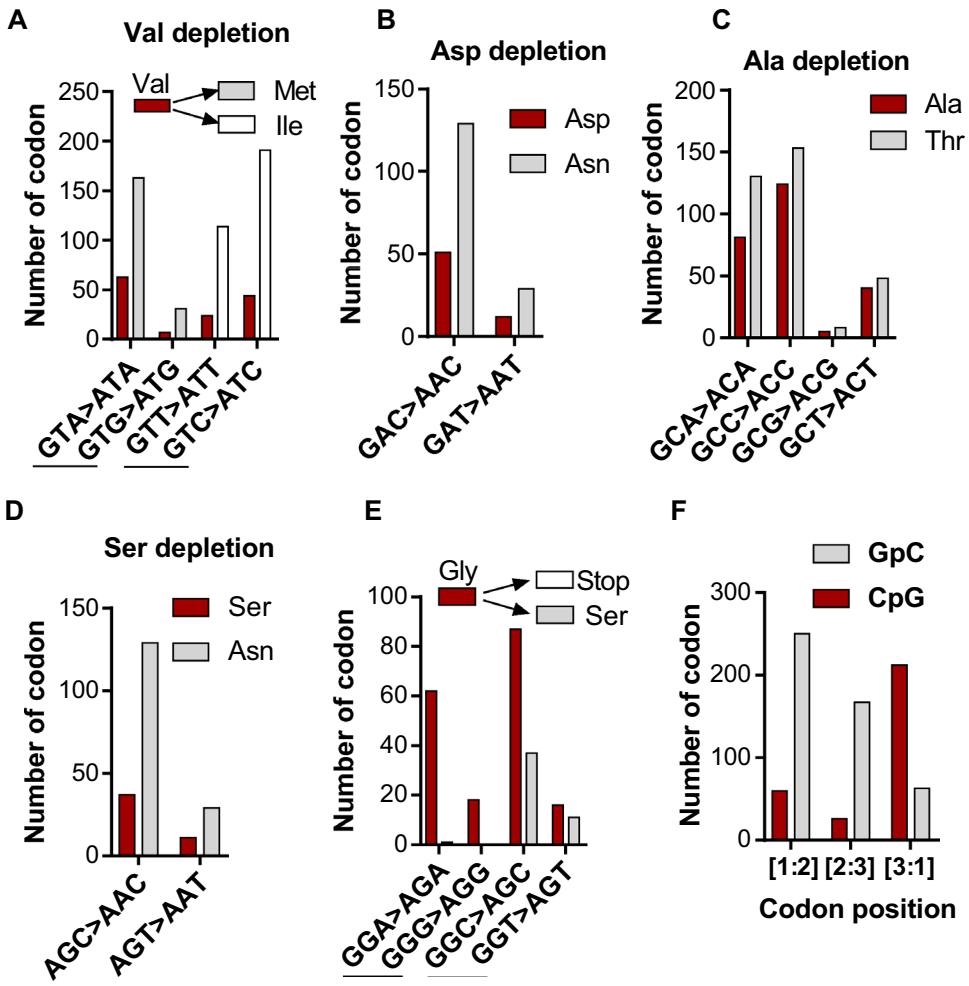


Fig. 5 Shaping of mitochondrial genome by mutation–selection interplay in germline mtDNA. Evolutionary depletion of codons in 12 L strand proteins for valine (**a**), aspartic acid (**b**), alanine (**c**) and serine (**d**) and the corresponding expansion of codons for methionine, isoleucine, asparagine, and threonine are shown, consistent with directional evolution. **e** The absence of glycine depletion is an example that directional evolution can be restrained by enough selection pressure. See also Fig. 3 for their respective mutation and selection pressure (marked by single asterisk (*). **f** Distribution of CpG and GpC sites at codon position [1:2], [2:3], and [3:1] in 12 L strand proteins. At codon site [1:2], CpG specifies arginine, while GpC specifies alanine. At codon site [2:3] and [3:1], the depletion of either CpG or GpC depends on whether G nucleotide sits at the neutral third codon position, but independent of the CpG or GpC dinucleotide motif

24 T > C, 13 A > G, 4 C > T). Taking the compositional bias of mtDNA into account, these pathogenic variants exhibited a 17.4-fold enrichment for G > A and a 2.4-fold enrichment for T > C, resembling the mutation signature of transition dominance and GT enrichment in germline mtDNA. Analysis of mutation occurrence revealed that several pathogenic mtDNA variants (G11778A, T14484C, A1555G, and G3460A) were mutational hotspots (Fig. 6a), accounting for their popularity in mitochondrial diseases (Chinnery 2015; Stewart and Chinnery 2015). Analysis of normalized GenBank frequency (corrected for mutation occurrence, Supplementary Table 3) for the 70 pathogenic transitions found a general correlation between variant frequency and disease severity, with variants causing severe mitochondrial diseases (MELAS and Leigh' syndrome) exhibiting lower frequency, and variants causing relatively mild disease (LHON, deafness) exhibiting higher frequency (Fig. 6b). These observations suggested that pathogenic mtDNA variants are also governed by the mutation–selection signature in germline mtDNA.

We further examined the distribution of normalized GenBank frequency for the 462 possibly pathogenic MITOMAP variants (160 G > A, 135 T > C, 112 A > G, and 55 C > T) using benign third codon variants as control and 7×10^{-5} as a tentative threshold (Fig. 6c). For benign third codon variants, biased distribution toward high frequency was most significant for G > A (binomial test $p = 2.2 \times 10^{-16}$), followed by T > C (binomial test $p < 0.05$), but not for AC targeting variants. However, for the 462 possibly pathogenic variants, a severely biased distribution toward low frequency was observed for G > A ($p < 2 \times 10^{-16}$), followed by T > C ($p = 4 \times 10^{-6}$), whereas AC targeting variants exhibited equal distribution between benign and possibly pathogenic variants (Fig. 6c). These data support that low frequency GT targeting variants, especially G > A transitions, are enriched for pathogenic variants.

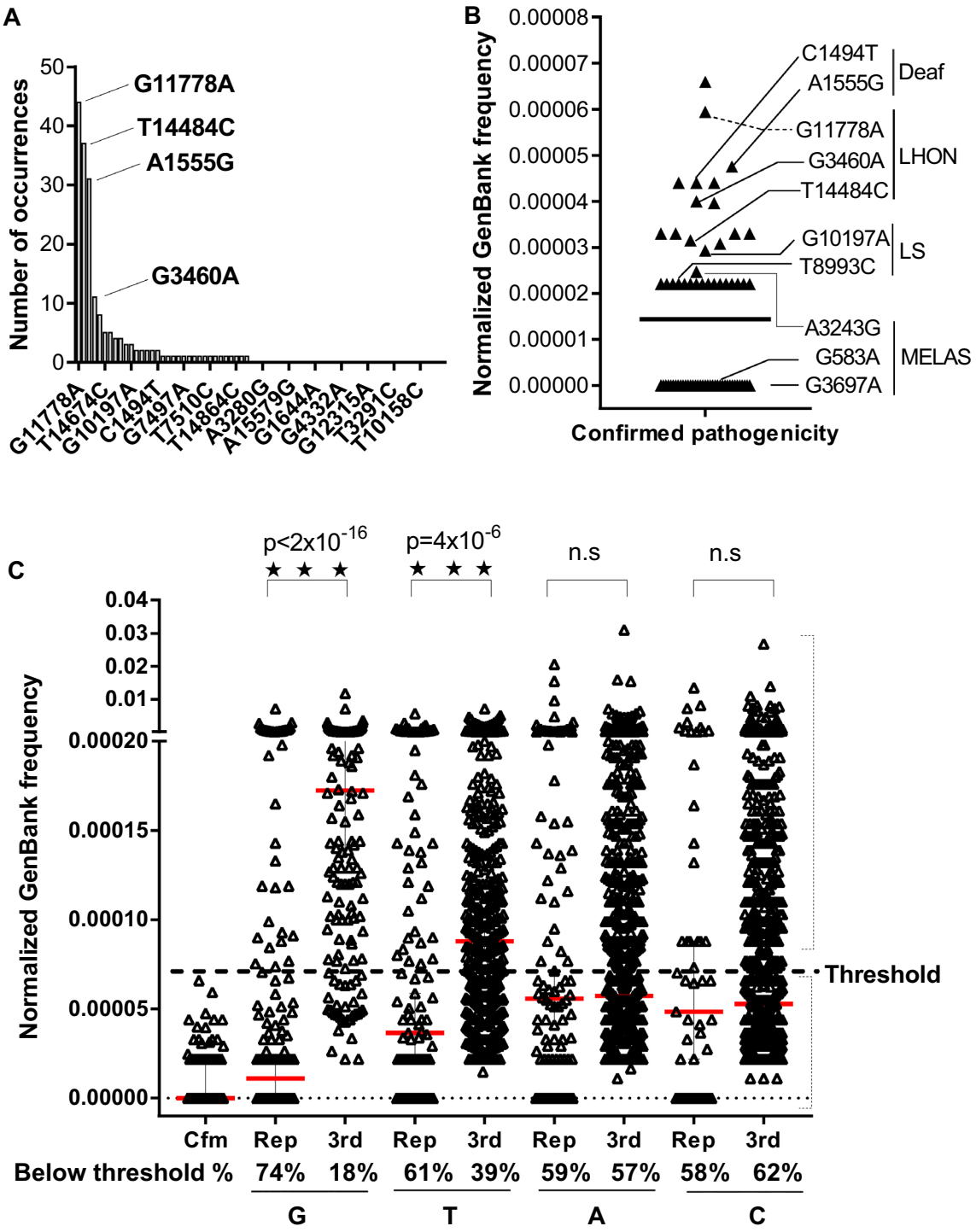


Fig. 6 Analysis of pathogenicity for disease-implicated mtDNA variants. **a** Number of mutation occurrence for 70 mtDNA transitions with confirmed pathogenicity based on combined historical and modern variants. The four mostly targeted variants (G11778A, T14484C, A1555G, and G3460A) are labeled. **b** Normalized GenBank frequency (corrected for number of occurrence) for 70 mtDNA transitions with confirmed pathogenicity. Representative variants for severe mitochondrial disease (MELAS, LS) and mild mitochondrial disease (LHON, deafness) are labeled. A correlation between normalized GenBank frequency and disease severity is evident. **c** Nucleotide-specific comparison of normalized GenBank frequency between 462 disease-implicated mtDNA transition variants and benign third codon variants. A threshold accommodating all the 70 confirmed pathogenic variants is shown by a dashed line. The percentage of variants below the threshold for each group is shown at the bottom. Nucleotide-specific significance between benign and disease-implicated variants was determined by Pearson's Chi squared test. Red line, median value of normalized frequency. Cfm, variants with confirmed pathogenicity. Rep, variants with reported but uncertain pathogenicity. 3rd, benign third codon variants. *n.s* not significant

Materials and methods

Extraction of historical and modern mitochondrial variants

We differentiated "historical" and "modern" mitochondrial variants, whereby "historical" variants represent the haplogroup defining variants, which probably reflect the historical mutational process, while "modern" variants represent rare variants in haplogroup, which likely reflect more recent mutational process. To harvest historical variants, we extracted the defining variants from the most updated Phylotree (Build 17, 18 Feb 2016) (van Oven and Kayser 2009). As these variants are phylogeny based, we assumed that each variant in the tree represents an independent mutation event. After removing indels and reversion mutations (mutations that reverse a mutated base back to its ancestral state), we gleaned a total of 11,746 "historical" mitochondrial variants. To harvest modern variants, we retrieved the most updated human mitochondrial variants from MITOMAP (www.mitomap.org), which contains a total of 97,566 variants from 45,494 full-length mtDNA sequences encompassing 1078 haplogroups. For each variant, we calculated a haplogroup-specific variant frequency (number of sequences carrying given variant in haplogroup/total number of sequences in the haplogroup) and defined variants with frequency of less than 5% as modern variants (each variant counted only once per haplogroup background). After removing indels and positions that are different between rCRS and RSRS reference (Behar et al. 2012), we gleaned a total of 44, 334 modern variants.

Analysis of mutational signature

Mutational signatures and strand bias were assessed following Alexandrov's approach (Alexandrov et al. 2013; Ju et al. 2014) with minor modifications. Briefly, because the initial analysis of mutation spectrum revealed predominance of transitions (96% for historical variants and 93% for modern variants), we only focused on the transition-associated mutation signatures, as the paucity of transversions did not allow for a solid analysis. We first picked the transition mutations and extracted the immediate 5′ and 3′ sequence context from mitochondrial RSRS. Transition rate was calculated as ratio between observed and expected mutation (H0 = equal mutation rate for all transition classes) for each of the 64 classes of trinucleotide context (16 types of possible 5′ and 3′ context × 4 classes of transition (A > G, C > T, G > A and T > C)). The trinucleotide frequency in human mtDNA reference was taken into account when calculating the expected mutation. We analyzed historical and modern variants independently in the first place, and combined them together when a great consistency between historical and modern variants was observed. For modern variants, 52 positions that differed between rCRS and RSRS reference were removed to avoid confusion. For analysis of strand bias, we transformed the purine transitions (G > A and A > G) into corresponding pyrimidine transition (C > T and T > C) on the complementary strand, and calculated the strand-specific C > T and T > C transition rate. The significance of strand bias (C > T and T > C, respectively) was determined by Welch's *t* test.

Functional unit-specific analysis of mutation signature

Historical and modern variants were divided into six functional units (tRNA, 12s rRNA, 16s rRNA, 1st, 2nd, and 3rd position of codon triplet) and independently analyzed for mutation signature and strand bias following the same approach. This functional dissection enabled the analysis of both mutation pressure and selection pressure.

For further analysis of neutral third codon variants, a gene-specific DssH value (the duration of parental H strand being single stranded during replication) was calculated following Reyes's approach (Faith and Pollock 2003; Reyes et al. 1998). Gene-specific G > A, T > C, A > G, and C > T transition rates were calculated by counting the number of specific third codon transitions in a given gene and further normalized for the third codon nucleotide compositions in the corresponding gene.

Analysis of selection pressure in the context of mutation pressure

We separated the codons of the 12 L strand proteins into 22 amino acid-based codon groups (two Leu codon, CTN and TTR, and two Ser codon, AGY and TCN, were treated as 4 different groups. R for purine, Y for pyrimidine and N for any of four nucleotides), and calculated codon-specific selection pressure according to the first and second codon position, respectively. The selection pressure was measured as the distance between the observed mutations and predicted selection-free mutational load. The selection-free mutations were estimated for each of the four nucleotides (A, C, G, and T) based on the mean observed mutation rate at the neutral third codon position. The H strand gene (MT-ND6) was not included for this analysis because (1) H and L strand are subjected to opposite mutation pressure in somatic cells (Ju et al. 2014) and (2) ND6 is a small protein with only 175 codons that did not allow for solid codon-specific analysis. We also calculated an overall region-specific selection pressure (tRNA, 12s RNA, 16s RNA, 1st 2nd and 3rd position of triplet codon, respectively) as the difference between observed and predicted selection-free mutational load. In addition, for the first, second, and third position in the codon triplet, we counted the overall percentage of

positions that are affected by mutation, serving as another measure of selection pressure.

Analysis of amino acid composition and CpG distribution

We counted the codon frequency in 12 mitochondrial L strand proteins based on the 22 amino acid-specific codon groups and evaluated whether the amino acid compositions can be explained by the mutation–selection interplay in germline mitochondrial genome. Another long-lasting mystery of mitochondrial genome is the pervasive depletion of CpG sites in animal mitochondrial genome (Cardon et al. 1994). We counted the distribution of CpG and GpC sites in 12 L strand proteins based on their relative position in the codon triplet ([1:2], [2:3], and [3:1], respectively) and evaluated whether the mutation signature in germline mtDNA can solve this conundrum.

Calculation of normalized GenBank frequency

To correct for the difference in mutation occurrence, we calculated a normalized GenBank frequency as $n/(m*45494)$, where n is the number of sequence harboring a given variant, m is the number of mutation incidence during evolution, and 45,494 is the total number of mtDNA sequence in MITOMAP. For parsimony, the number of incidences for both historical and modern variants was combined. Pearson's correlation was performed to evaluate the consistency of mutation occurrence between historical and modern variants.

Analysis of mitochondrial variants with reported pathogenicity

We retrieved 532 mtDNA transition variants located within the coding region from a total of 680 disease-implicated variants from MITOMAP. Among them, only 70 variants had confirmed pathogenicity ("Cfm" status), while the remaining 462 variants were reported to be possibly pathogenic but needed further validation ("Reported" "Reported secondary", or "Unclear" status). We applied the calculated normalized GenBank frequency for these disease-implicated variants and evaluated whether they could be used to predict pathogenicity by using the 70 pathogenic variants as positive control and the neutral third codon position as negative control.

Statistical analysis

Data processing and statistical testing were performed using R software. All p values were calculated by two-tailed testing. Figures were generated using GraphPad prism software.

Acknowledgements We are grateful to Prof. Tielin Yang, Prof. Xiaogang Liu (Xi'an Jiaotong University), Prof. Jianhua Zheng (Zhengzhou University), and Prof. Douglas C. Wallace (Children's Hospital of Philadelphia) for helpful discussions and critical reading of the manuscript. This work was supported by the Fundamental Research Funds for the Central Universities (to XG), the Scientific Research Foundation for Returned Scholars of Shaanxi Province (to XG), the National Basic Research Program (973 Project 2015CB553602 to JL), and the National Natural Science Foundation of China (91649106, 31770917, 31570777 to JL).

Author contributions XG conceived the idea, designed the research, performed the analysis, analyzed data, and wrote the paper; XK analyzed the data, JL analyzed the data and co-wrote the paper. All authors reviewed the manuscript.

Compliance with ethical standards

Conflict of interest The authors declare no competing financial interests.

References

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjord JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Jager N, Jones DT, Jones D, Knappskog S, Kool M, Lakhani SR, Lopez-Otin C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt AN, Valdes-Mas R, van Buuren MM, van Veer L, Vincent-Salomon A, Waddell N, Yates LR, Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, Stratton MR (2013) Signatures of mutational processes in human cancer. Nature 500:415–421. https://doi.org/10.1038/nature12477

Behar DM, van Oven M, Rosset S, Metspalu M, Loogvali EL, Silva NM, Kivisild T, Torroni A, Villems R (2012) A "Copernican" reassessment of the human mitochondrial DNA tree from its root. Am J Hum Genet 90:675–684. https://doi.org/10.1016/j.ajhg.2012.03.002

Cardon LR, Burge C, Clayton DA, Karlin S (1994) Pervasive CpG suppression in animal mitochondrial genomes. Proc Natl Acad Sci 91:3799–3803

Chinnery PF (2015) Mitochondrial disease in adults: what's old and what's new? EMBO Mol Med 7:1503–1512. https://doi.org/10.15252/emmm.201505079

Faith JJ, Pollock DD (2003) Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. Genetics 165:735–745

Fan W, Waymire KG, Narula N, Li P, Rocher C, Coskun PE, Vannan MA, Narula J, Macgregor GR, Wallace DC (2008) A mouse model of mitochondrial disease reveals germline selection against severe mtDNA mutations. Science 319:958–962. https://doi.org/10.1126/science.1147786

Floros VI, Pyle A, Dietmann S, Wei W, Tang WCW, Irie N, Payne B, Capalbo A, Noli L, Coxhead J, Hudson G, Crosier M, Strahl H, Khalaf Y, Saitou M, Ilic D, Surani MA, Chinnery PF (2018) Segregation of mitochondrial DNA heteroplasmy through a

developmental genetic bottleneck in human embryos. Nat Cell Biol 20:144–151. https://doi.org/10.1038/s41556-017-0017-8

Fonseca MM, Harris DJ, Posada D (2014) The inversion of the control region in three mitogenomes provides further evidence for an asymmetric model of vertebrate mtDNA replication. PLoS One 9:e106654. https://doi.org/10.1371/journal.pone.0106654

Helleday T, Eshtad S, Nik-Zainal S (2014) Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet 15:585–598. https://doi.org/10.1038/nrg3729

Itsara LS, Kennedy SR, Fox EJ, Yu S, Hewitt JJ, Sanchez-Contreras M, Cardozo-Pelaez F, Pallanck LJ (2014) Oxidative stress is not a major contributor to somatic mitochondrial DNA mutations. PLoS Genet 10:e1003974. https://doi.org/10.1371/journal.pgen.1003974

Ju YS, Alexandrov LB, Gerstung M, Martincorena I, Nik-Zainal S, Ramakrishna M, Davies HR, Papaemmanuil E, Gundem G, Shlien A, Bolli N, Behjati S, Tarpey PS, Nangalia J, Massie CE, Butler AP, Teague JW, Vassiliou GS, Green AR, Du MQ, Unnikrishnan A, Pimanda JE, Teh BT, Munshi N, Greaves M, Vyas P, El-Naggar AK, Santarius T, Collins VP, Grundy R, Taylor JA, Hayes DN, Malkin D, Foster CS, Warren AY, Whitaker HC, Brewer D, Eeles R, Cooper C, Neal D, Visakorpi T, Isaacs WB, Bova GS, Flanagan AM, Futreal PA, Lynch AG, Chinnery PF, McDermott U, Stratton MR, Campbell PJ (2014) Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. Elife. https://doi.org/10.7554/elife.02935

Kauppila JH, Stewart JB (2015) Mitochondrial DNA: radically free of free-radical driven mutations. Biochim Biophys Acta 1847:1354–1361. https://doi.org/10.1016/j.bbabio.2015.06.001

Kennedy SR, Salk JJ, Schmitt MW, Loeb LA (2013) Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. PLoS Genet 9:e1003794. https://doi.org/10.1371/journal.pgen.1003794

Kivisild T, Shen P, Wall DP, Do B, Sung R, Davis K, Passarino G, Underhill PA, Scharfe C, Torroni A, Scozzari R, Modiano D, Coppa A, de Knijff P, Feldman M, Cavalli-Sforza LL, Oefner PJ (2006) The role of selection in the evolution of human mitochondrial genomes. Genetics 172:373–387. https://doi.org/10.1534/genetics.105.043901

Kloss-Brandstatter A, Pacher D, Schonherr S, Weissensteiner H, Binna R, Specht G, Kronenberg F (2011) HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum Mutat 32:25–32. https://doi.org/10.1002/humu.21382

Lightowlers RN, Taylor RW, Turnbull DM (2015) Mutations causing mitochondrial disease: what is new and what challenges remain? Science 349:1494–1499. https://doi.org/10.1126/science.aac7516

Lott MT, Leipzig JN, Derbeneva O, Xie HM, Chalkia D, Sarmady M, Procaccio V, Wallace DC (2013) mtDNA variation and analysis using mitomap and mitomaster. Curr Protoc Bioinform. https://doi.org/10.1002/0471250953.bi0123s44

Nicholls TJ, Minczuk M (2014) In D-loop: 40 years of mitochondrial 7S DNA. Exp Gerontol 56:175–181. https://doi.org/10.1016/j.exger.2014.03.027

Pereira L, Soares P, Radivojac P, Li B, Samuels DC (2011) Comparing phylogeny and the predicted pathogenicity of protein variations reveals equal purifying selection across the global human mtDNA diversity. Am J Hum Genet 88:433–439. https://doi.org/10.1016/j.ajhg.2011.03.006

Raina SZ, Faith JJ, Disotell TR, Seligmann H, Stewart CB, Pollock DD (2005) Evolution of base-substitution gradients in primate mitochondrial genomes. Genome Res 15:665–673. https://doi.org/10.1101/gr.3128605

Rebolledo-Jaramillo B, Su MS, Stoler N, McElhoe JA, Dickins B, Blankenberg D, Korneliussen TS, Chiaromonte F, Nielsen R, Holland MM, Paul IM, Nekrutenko A, Makova KD (2014) Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA. Proc Natl Acad Sci USA 111:15474–15479. https://doi.org/10.1073/pnas.1409328111

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

Sonney S, Leipzig J, Lott MT, Zhang S, Procaccio V, Wallace DC, Sondheimer N (2017) Predicting the pathogenicity of novel variants in mitochondrial tRNA with MitoTIP. PLoS Comput Biol 13:e1005867. https://doi.org/10.1371/journal.pcbi.1005867

Stewart JB, Chinnery PF (2015) The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. Nat Rev Genet 16:530–542. https://doi.org/10.1038/nrg3966

Stewart JB, Freyer C, Elson JL, Larsson NG (2008a) Purifying selection of mtDNA and its implications for understanding evolution and mitochondrial disease. Nat Rev Genet 9:657–662. https://doi.org/10.1038/nrg2396

Stewart JB, Freyer C, Elson JL, Wredenberg A, Cansu Z, Trifunovic A, Larsson NG (2008b) Strong purifying selection in transmission of mammalian mitochondrial DNA. PLoS Biol 6:e10. https://doi.org/10.1371/journal.pbio.0060010

Stewart JB, Alaei-Mahabadi B, Sabarinathan R, Samuelsson T, Gorodkin J, Gustafsson CM, Larsson E (2015) Simultaneous DNA and RNA mapping of somatic mitochondrial mutations across diverse human cancers. PLoS Genet 11:e1005333. https://doi.org/10.1371/journal.pgen.1005333

van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30:E386–E394. https://doi.org/10.1002/humu.20921

Vyas S, Zaganjor E, Haigis MC (2016) Mitochondria and cancer. Cell 166:555–566. https://doi.org/10.1016/j.cell.2016.07.002

Wallace DC, Chalkia D (2013) Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. Cold Spring Harb Perspect Biol 5:a021220. https://doi.org/10.1101/cshperspect.a021220

Wanrooij S, Falkenberg M (2010) The human mitochondrial replication fork in health and disease. Biochim Biophys Acta 1797:1378–1388

Williams SL, Mash DC, Zuchner S, Moraes CT (2013) Somatic mtDNA mutation spectra in the aging human putamen. PLoS Genet 9:e1003990. https://doi.org/10.1371/journal.pgen.1003990

Zeng AGX, Leung ACY, Brooks-Wilson AR (2018) Somatic mitochondrial DNA mutations in diffuse large B-cell lymphoma. Sci Rep 8:3623. https://doi.org/10.1038/s41598-018-21844-6