Materials and Methods

Reference genomes and annotations. The reference sequence for the human nuclear genome was GRCh37/hg19, as downloaded from the 1000 Genomes Project data server (http://www.1000genomes.org/). The revised Cambridge Reference Sequence (rCRS) and gene annotations for the human mitochondrial genome were downloaded from NCBI with accession number NC_012920. So were the reference mitochondrial genomes and annotations for *Pan troglodytes*, and *Pongo abelii*. Annotation of synonymous and non-synonymous changes for rCRS, and the secondary structure of tRNA and rRNA was retrieved from a previous study (1). The secondary structure of tRNA and rRNA were computed with the mfold program (2). Relative Mutation Rate (RMR) for each site was inferred as the absolute frequency of occurrence of the mutation in a phylogenetic tree constructed with 2196 global human samples (3).

Sequencing data. Sequencing reads mapped to the mitochondrial genome in the 1000 Genomes Project phase 1 data were downloaded from the 1000 genomes data server. Our analysis focused on 1085 unrelated individuals from 14 populations, which were sequenced using either ILLUMINA or SOLID platforms. There were 9 individuals sequenced by two methods (ILLUMINA and LS454). These individuals were used to confirm the reliability of our computational pipeline with ILLUMINA data. See Table S1 for more detailed information.

Definition of ancestral alleles. A previously described method was used to define ancestral human mtDNA alleles with high confidence (4). First, LASTZ (5) was used to align the mitochondrial genomes of *Homo sapiens*, *Pan troglodytes*, and *Pongo abelii*. Furthermore, to take advantage of the better conservativeness of protein sequences than DNA sequences, we aligned the coding region based on MUSCLE alignments of protein sequences (6). Only alleles that were consistent in both *Pan troglodytes* and *Pongo abelii*, and also present in *Homo sapiens* were considered as the ancestral alleles.

Computational pipeline for calling heteroplasmy and polymorphism. Sequencing reads retrieved from the 1000 genome data sever were re-mapped to the combined human genome, both nuclear and mitochondrial genomes, using GSNAP (7). Following previous practice (8), we counted unknown characters (N) as mismatches (--query-unk-mismatch=1) and only retained sequences that mapping uniquely to the genome (-n 1 -Q). Another important parameter for mapping is the maximum number of mismatches allowed (-m). By default, the parameter is ((readlength+2)/15 - 2), corresponding to 5 mismatches for read length of 100bp. In our analysis, using the default parameters resulted in unsatisfactory coverage, especially for non-European individuals. This is due to the fact that mitochondrial DNA is much more divergent than nuclear DNA (1), and the reference mitochondrial DNA is from an individual of European origin (9). To accommodate this fact, we adjusted the parameter to allow 7% mismatches (-m 0.07), corresponding to 7 mismatches for a read length of 100 bp. To confirm that our observed patterns are not artifacts of mis-mapping, we applied both the default and the adjusted parameters. Both parameters yielded similar patterns of heteroplasmy. Only results using a 7% mismatch threshold are presented.

After the GSNAP reads mapping, we recorded only reads that are uniquely mapped to the mitochondrial genome in order to minimize the complications of nuclear mitochondrial

sequences (NumtS) (10). We further filtered the data and defined "usable sites" based on the following three quality control criteria: 1) only bases with Phred quality score >= 20 were used; 2) only sites with 10X coverage of qualified bases on both positive and negative strands were used; 3) only sites that satisfy criteria 1) and 2) in more than 95% individuals were used in analysis of heteroplasmy and polymorphism. A candidate heteroplasmic site was defined with the following two criteria: 1) the raw frequency for the minor allele is no less than 1% on both strands; 2) all alleles have support from at least 2 reads on each strand.

For each candidate heteroplasmic site, we further applied a maximum likelihood (ML) method to accurately estimate the frequency of the major allele while taking into account sequencing error (8, 11). For example, for all bases mapped to the positive strand of a locus, l bases are the major alleles and k bases are the minor alleles. Each base has respective sequencing quality, corresponding to the probability of sequencing error ε. The underlying parameter of interest is the frequency of the major allele f. The likelihood function could be written as follows:

$$L(f) = \prod_{j=1}^{l}\left[(1-f)\varepsilon_j + f\left(1-\varepsilon_j\right)\right]\prod_{j=1}^{k}\left[(1-f)(1-\varepsilon_j) + f\varepsilon_j\right]$$

We estimated f under two models: heteroplasmy (m_1) and homoplasmy (m_0). And a log-likelihood ratio (LLR) was calculated as $log\left(\frac{L(\hat{f}_{m_1})}{L(\hat{f}_{m_0})}\right)$. A high-confidence heteroplasmy was defined as candidate heteroplasmy with LLR no less than 5 (8). With all these criteria (See Table S2 for a brief list), a total of 4342 heteroplasmies were defined. Among them, 153 have a minor allele frequency estimated by the ML method to be smaller than 1%, even though we required that the raw frequency for the minor allele is no less than 1% on both strands.

After detecting heteroplasmy, consensus sequences were assembled for each individual and compared among all individuals to identify polymorphic sites. Only "usable sites" satisfying the above-mentioned criteria were considered. For each individual, a consensus sequence was assembled using the alleles present at homoplasmic sites, and the major alleles at heteroplasmic sites. Sites were classified as polymorphic if there was more than one allele present in the consensus sequences of the population.

To confirm the reliability of our computational pipeline in defining heteroplasmy, we took advantage of the 9 individuals sequenced by both ILLUMINA and LS454. LS454 data were directly retrieved from the 1000 genome data sever and processed as followed: 1) Only loci defined as heteroplasmy in ILLUMINA data were examined; 2) Only reads with mapping quality no less than 20 and bases with sequencing quality no less than 20 were used; 3) Assuming a biallelic state, only the two most common alleles were retained; 4) The frequency of the heteroplasmic alleles were estimated with the ML method described above. Only heteroplasmy with the same alleles as identified by ILLUMINA was considered as confirmed.

The measure of pathogenicity. The pathogenicity scores for all possible non-synonymous changes were retrieved from a previous study (12). All possible non-synonymous changes were

inferred based on the rCRS sequence and the pathogenicity of a non-synonymous change was predicted with the MutPred algorithm (13). A higher pathogenicity score indicates a higher likelihood that the non-synonymous change is pathogenic. Three types of attributes were utilized by MutPred in classifying amino acid variations: 1) attributes based on predicted protein structure and dynamics including secondary structure, solvent accessibility, transmembrane helices, coiled-coil structure, stability, B-factor, and intrinsic disorder; 2) attributes based on predicted functional properties such as DNA-binding residuals, catalytic residues, calmodulin-binding targets, and sites of phosphorylation, methylation, ubiquitination and glycosylation; 3) attributes based on amino acid sequence and evolutionary information, including sequence conservativeness, SIFT score, Pfam profile score, and transition frequencies. The software is trained with a random forest classification model to discriminate between disease-associated amino acid substitution from the Human Gene Mutation Database and putatively neutral polymorphisms from Swiss-Prot (12, 13).

The pathogenic effect of all possible non-synonymous changes were also predicted by PolyPhen-2 (14, 15). PolyPhen-2 combines sequence- and structure-based attributes and predicts the effect of missense mutation with a naive Bayesian classifier. The default HumDiv-trained predictor was used in this study. The pathogenicity predicted by MutPred and Polyphen is highly consistent (Fig. S8).

The pathogenic effect of tRNA mutations were downloaded from a previous publication (16). A tRNA mutation was deemed deleterious by a computational method taking into account the following attributes: 1) evolutionary conservation; 2) disruption of Watson-Crick pairing; 3) the tendency of co-evolution by complementary mutation in the stem.

Disease association information was obtained from MITOMAP (17).

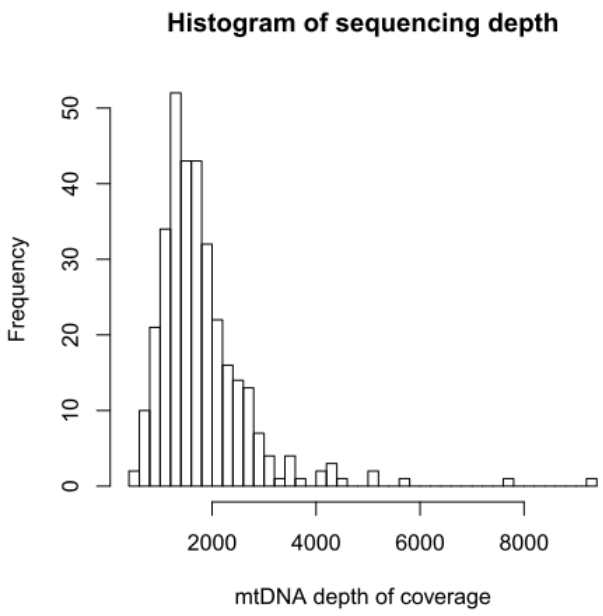


Fig. S1. The histogram of sequencing depth for mtDNA in 1085 individuals.

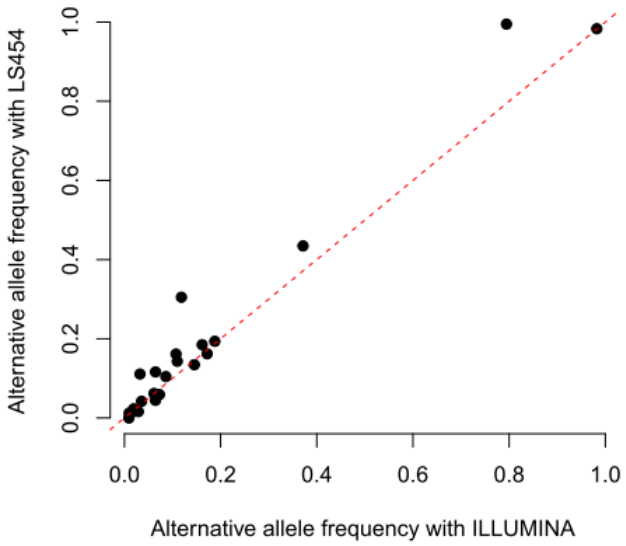


Fig. S2. The comparison of alternative allele frequencies for heteroplasmies identified in 9 individuals sequenced by both ILLUMINA and LS454. The allele frequencies were estimated by ML method.

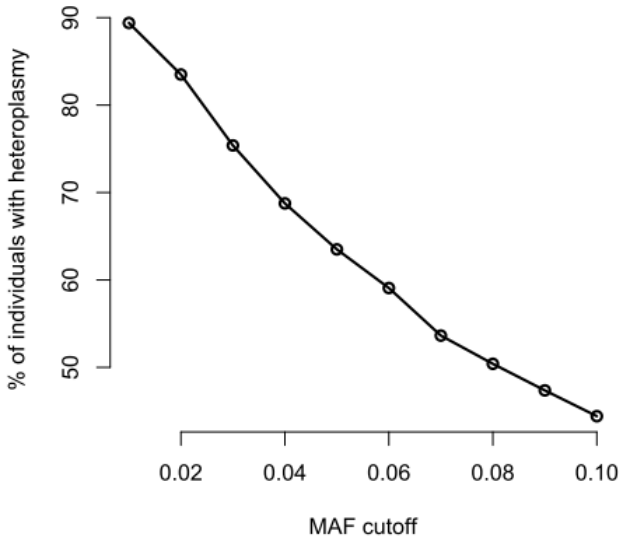


Fig. S3. The prevalence of heteroplasmy in the sample with different MAF cutoff in definition of heteroplasmy.

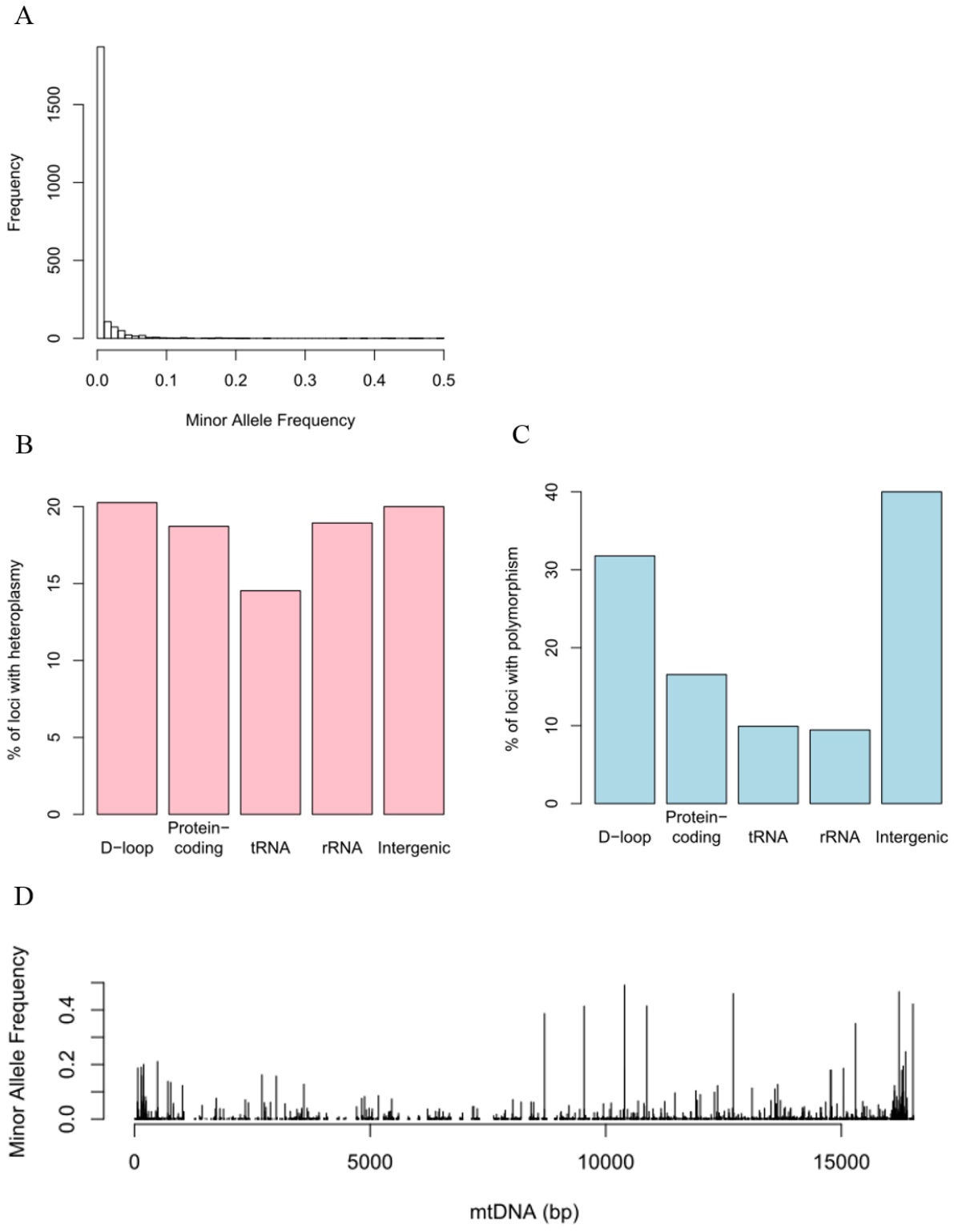


Fig. S4. The prevalence of heteroplasmy and polymorphisms in mtDNA. **A.** The histogram for minor allele frequency of polymorphism. **B.** The prevalence of heteroplasmy in each genomic

region; **C.** The prevalence of polymorphism in each genomic region; **D.** The genomic distribution of polymorphisms and their minor allele frequency in the sample of 1085 individuals.

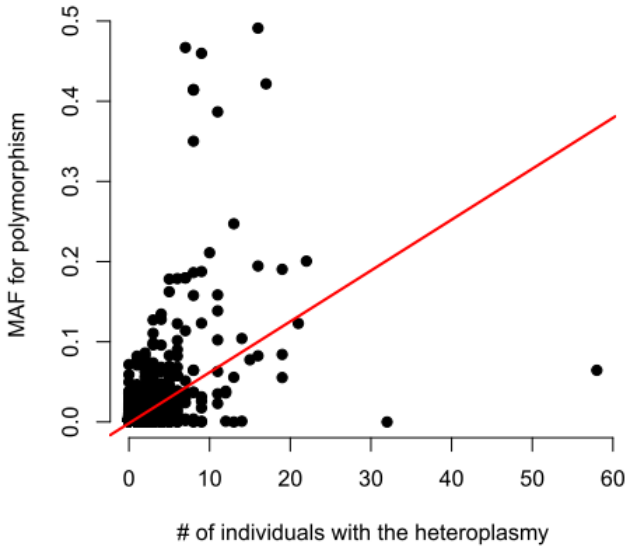


Fig. S5. The positive correlation between the incidence of heteroplasmy and the minor allele frequency of polymorphism in the sample of 1085 individuals. Each dot represent a locus that is polymorphic or heteroplasmic.

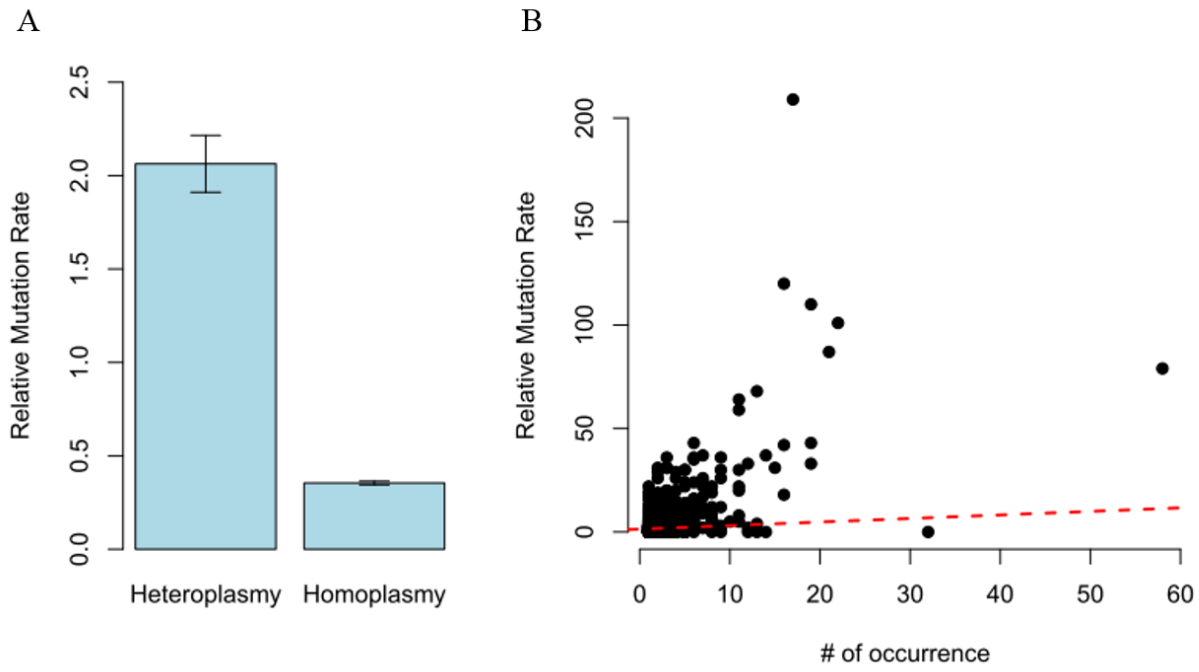


Fig. S6. Mutation rate in mtDNA and heteroplasmy. A. The barplot of relative mutation rate for heteroplasmic and homoplasmic loci. Error bar represents one standard error. **B.** The positive correlation between relative mutation rate and the number of occurrence in the population. Each black dot represents a heteroplasmic locus. And the red dashed line indicates the linear regression.

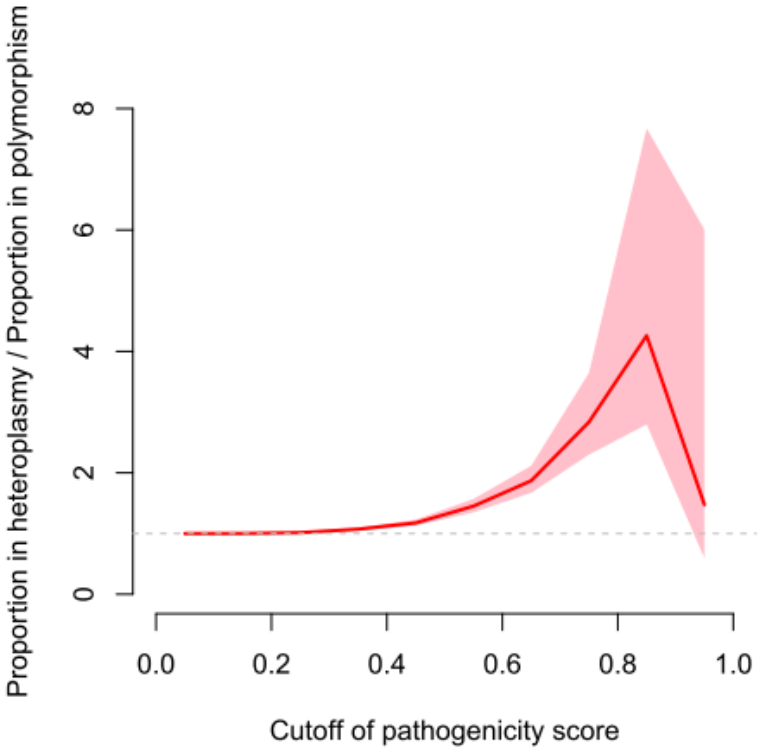


Fig. S7. The relative risk of heteroplasmy being pathogenic when compared with polymorphism. A pathogenic mutation is defined with varying cutoff of pathogenicity score. The red line is the empirical observation while the pink region represent the 95% bootstrap confidence interval.

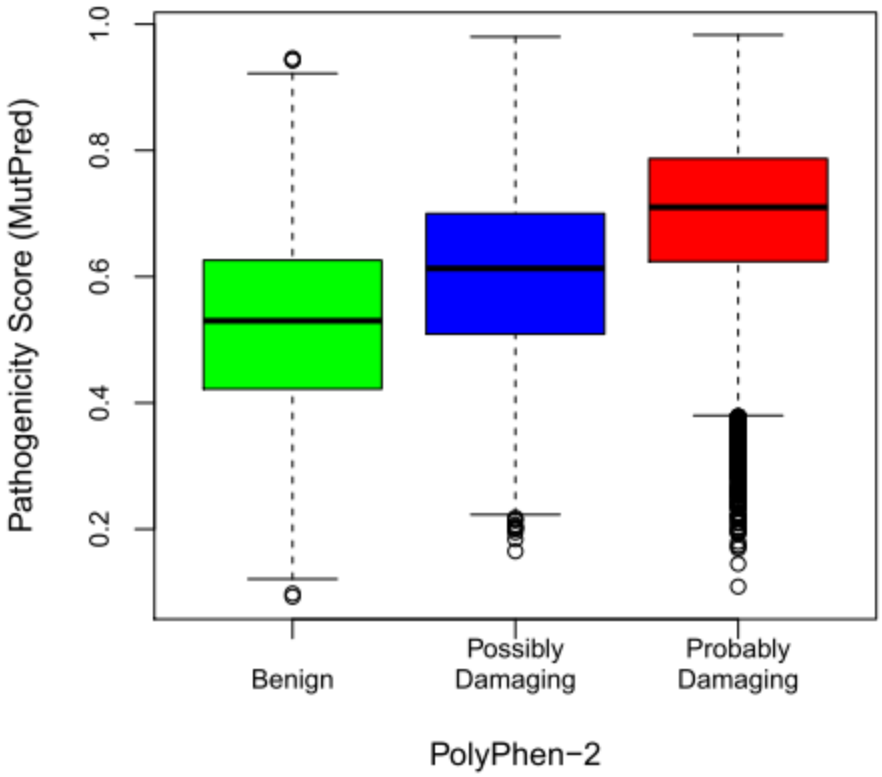


Fig. S8. Consistent pathogenicity as predicted by MutPred and PolyPhen-2. The MutPred pathogenicity scores for the three functional categories predicted by PolyPhen-2.

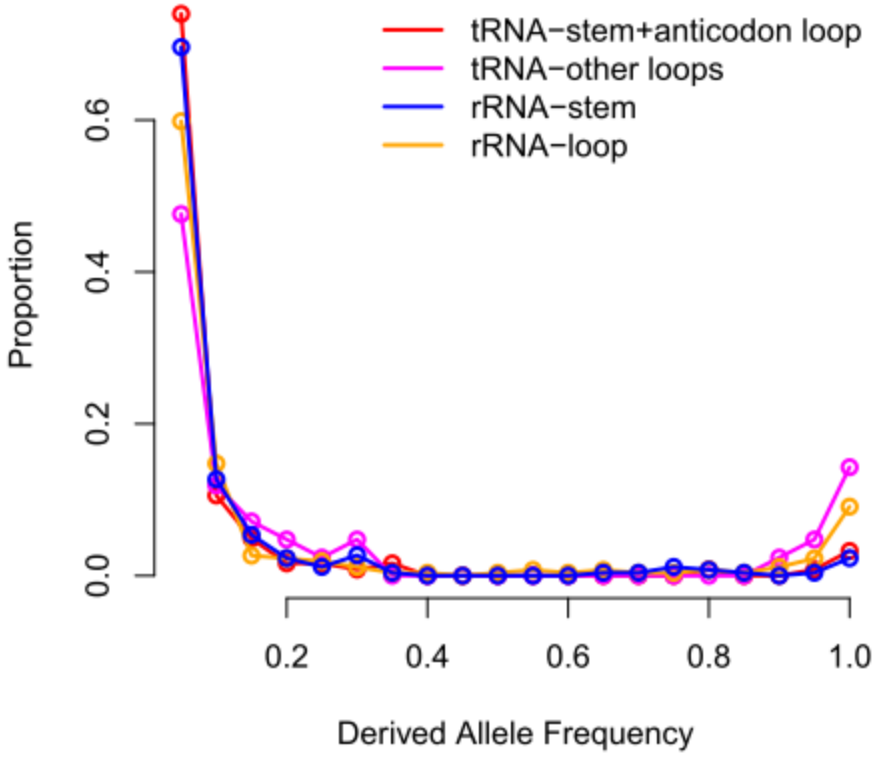


Fig. S9. The distribution of derived allele frequency for heteroplasmies in different regions in tRNA and rRNA.

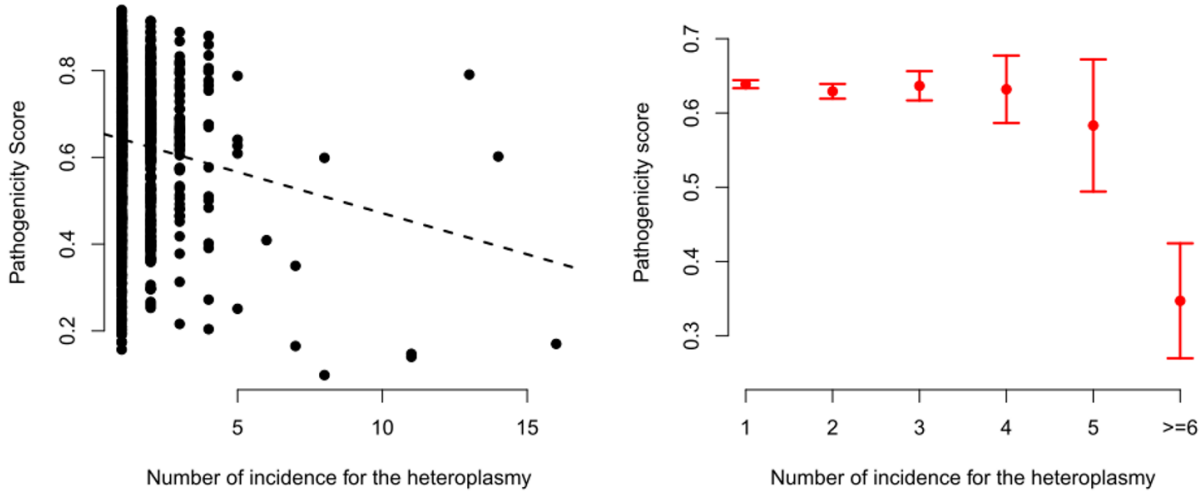
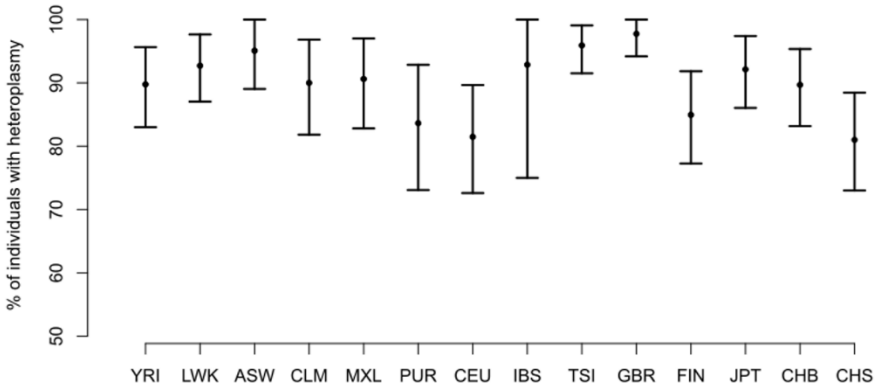
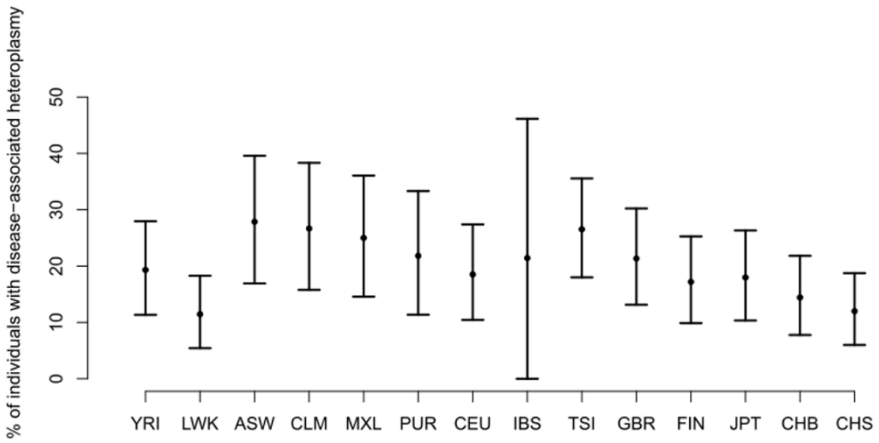


Fig. S10. The negative relationship between pathogenicity score and the number of incidence of heteroplasmy in the population. A. Each dot represents one heteroplasmy. **B.** Similar presentation with A where heteroplasmies are binned based on their incidence.

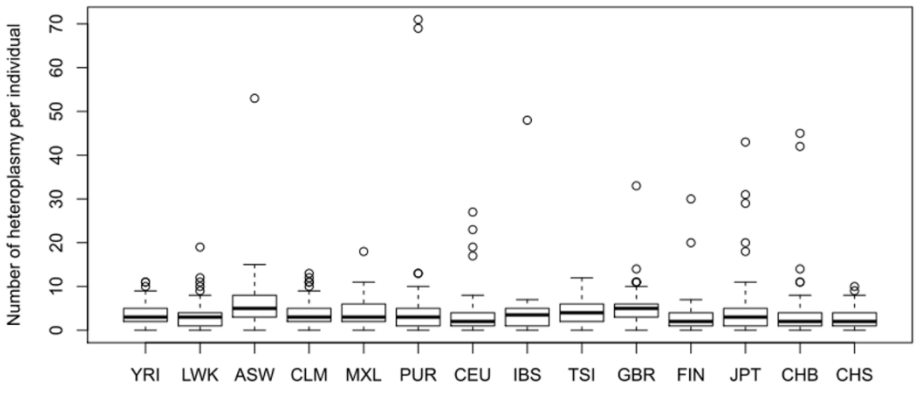
A


% of individuals with heteroplasmy
50
60
70
80
90
100
YRI
LWK
ASW
CLM
MXL
PUR
CEU
IBS
TSI
GBR
FIN
JPT
CHB
CHS

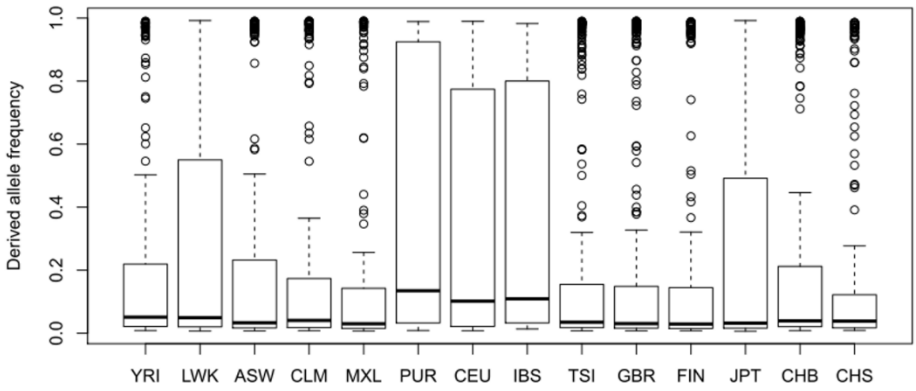
B


% of individuals with disease−associated heteroplasmy
0
10
20
30
40
50
YRI
LWK
ASW
CLM
MXL
PUR
CEU
IBS
TSI
GBR
FIN
JPT
CHB
CHS

C


Number of heteroplasmy per individual
0
10
20
30
40
50
60
70
YRI
LWK
ASW
CLM
MXL
PUR
CEU
IBS
TSI
GBR
FIN
JPT
CHB
CHS

D



E

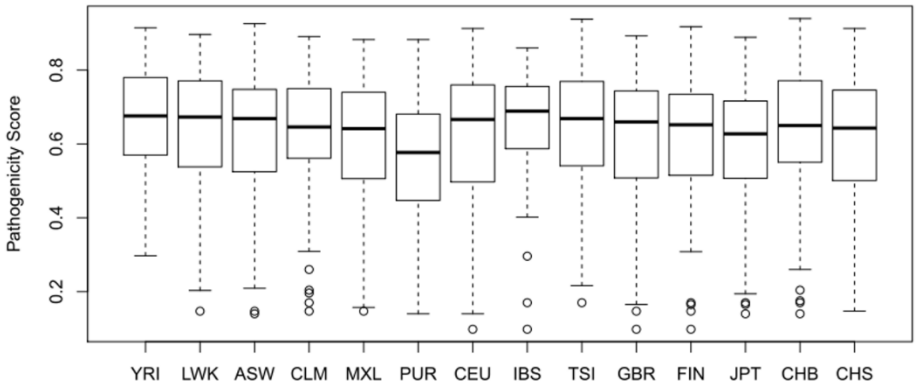


Fig. S11. Similar heteroplasmy pattern across different human populations. Inter-population comparisons of: **A.** the percentage of individuals carrying at least one heteroplasmy; **B.** the percentage of individuals carrying at least one disease-associated heteroplasmy; **C.** the number of heteroplasmy per individual; **D.** derived allele frequency; **E.** pathogenicity score of non-synonymous heteroplasmies. The error bars in A and B represent 95% CI from 10^5 bootstraps of individuals. For A and B, pairwise comparisons were performed with permutation test. For C. D and E, pairwise comparison were performed with Wilcoxon rank-sum test. Bonferroni corrections were performed with 92 tests including the comparison of male and female. None of the population achieve significance in all comparisons with other populations.

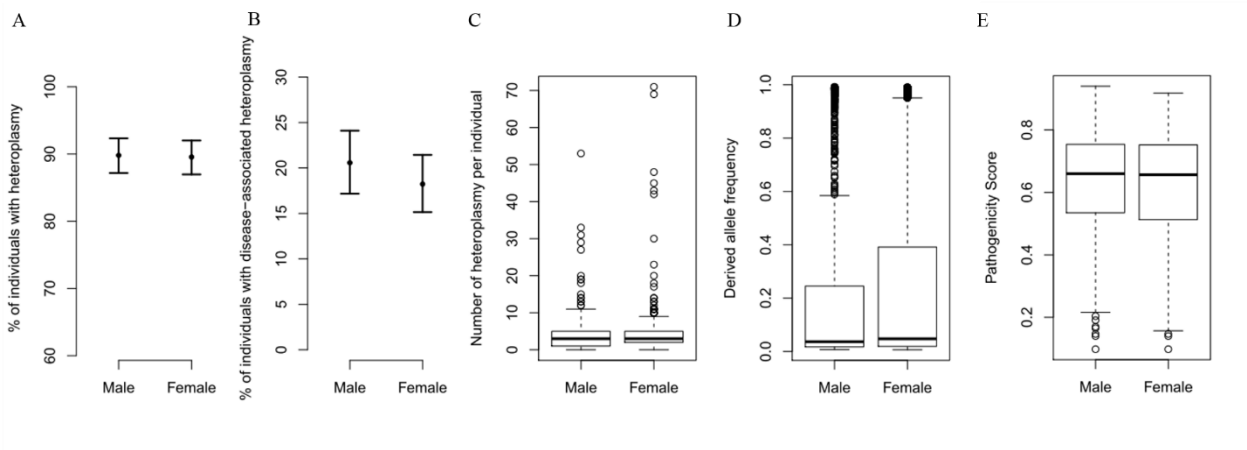


Fig. S12. Similar heteroplasmy pattern between genders. Inter-gender comparisons of: **A.** the percentage of individuals carrying at least one heteroplasmy; **B.** the percentage of individuals carrying at least one disease-associated heteroplasmy; **C.** the number of heteroplasmy per individual; **D.** derived allele frequency; **E.** pathogenicity score of non-synonymous heteroplasmies. The error bars in A and B represent 95% CI from 10^5 bootstraps of individuals. For A and B, pairwise comparisons were performed with permutation test. For C. D and E, pairwise comparison were performed with Wilcoxon rank-sum test. Bonferroni corrections were performed with 92 tests including the inter-population comparisons. No significance were found after Bonferroni correction.

Table S1. Sequencing data from 1000 genome project

Population	# by ILLUMINA	# by SOLID	Total
ASW	50	11	61
CEU	81	0	81
CHB	81	16	97
CHS	92	8	100
CLM	50	10	60
FIN	75	18	93
GBR	70	19	89
IBS	6	8	14
JPT	78	11	89
LWK	82	14	96
MXL	52	12	64
PUR	52	3	55
TSI	98	0	98
YRI	76	12	88

ASW: Americans of African Ancestry in SW USA; CEU: Utah Residents (CEPH) with Northern and Western European ancestry; CHB: Han Chinese in Bejing, China; CHS: Southern Han Chinese; CLM: Colombians from Medellin, Colombia; FIN: Finnish in Finland; GBR: British in England and Scotland; IBS: Iberian population in Spain; JPT: Japanese in Tokyo, Japan; LWK: Luhya in Webuye, Kenya; MXL: Mexican Ancestry from Los Angeles USA; PUR: Puerto Ricans from Puerto Rico; TSI: Toscani in Italia; YRI: Yoruba in Ibadan, Nigera. Populations highlighted in blue are those of European ancestry.

Table S2. Comparison of criteria for calling heteroplasmy.

	He et al. 2010 Nature(18)	Li et al. 2010, AJHG(19)	Goto et al. 2011, Genome Biology(20)	Picardi & Pesole, 2012, Nature Methods(8)	This study
Sequencing Technology	ILLUMINA ~16,700X	ILLUMINA 36 & 76 bp ~67 & ~211X	ILLUMINA ~1,170X	ILLUMINA, Agilent, NimbleGen	ILLUMINA, SOLID
Mapping Tools	Eland	MIA	BWA	GSNAP	GSNAP remapping
Mismatches	<= 3 in 36 bp	Default	Default	Default	Default or 7%
Reads	remove low-quality reads	remove duplicate reads & low-quality reads	--	--	Only reads originally mapped to mtDNA by 1000 genome project
Mapping	--	--	Unique	Unique	Unique
Base Quality	>=23 for all bases in the read	>= 20 on site; >= 15 for 5 bp flanking	>= 30 on site	>=20 on site	>=20 on site
Minimum Depth	>= 10 distinct reads	--	>= 100X HQ depth on each strand	>=20 X >=2 reads	>= 10X HQ depth on each strand
Double Strand Validation (control for context-dependent error & PCR duplicate during sequencing)	>=3 reads on each strand	>= 1 read on each strand	>=100 HQ depth on each strand; >= 2% raw frequency[a] on each strand;	--	>= 2 reads on each strand; >=1% raw frequency on each strand
Minor allele frequency[b]	>=1.6%	>=10%	>= 2% on each strand	--	>=1% on each strand
Log-likelihood ratio	--	--	--	>=5	>=5

a. Raw frequency for each locus was calculated as the fraction of the allele among all observed alleles. This is in contrast to frequency estimated with maximum likelihood method which takes into account sequencing error.

The minor allele frequency used in all studies are based on raw frequency.

References:

1. Pereira L, *et al.* (2009) The diversity present in 5140 human mitochondrial genomes. *Am J Hum Genet* 84(5): 628-640.
2. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31(13): 3406-3415.
3. Soares P, *et al.* (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet* 84(6): 740-759.
4. Keinan A, Mullikin JC, Patterson N, Reich D (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet* 39(10): 1251-1255.
5. Harris RS (2007) Improved pairwise alignment of genomic DNA. (The Pennsylvania State University).
6. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5): 1792-1797.
7. Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26(7): 873-881.
8. Picardi E, Pesole G (2012) Mitochondrial genomes gleaned from human whole-exome sequencing. *Nat Methods* 9(6): 523-524.
9. Andrews RM, *et al.* (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* 23(2): 147.
10. Simone D, *et al.* (2011) The reference human nuclear mitochondrial sequences compilation validated and implemented on the UCSC genome browser. *BMC Genomics* 12: 517.
11. Chepelev I (2012) Detection of RNA editing events in human cells using high-throughput sequencing. *Methods Mol Biol* 815: 91-102.
12. Pereira L, *et al.* (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(4): 433-439.
13. Li B, *et al.* (2009) Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* 25(21): 2744-2750.
14. Adzhubei IA, *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat Methods* 7(4): 248-249.
15. Adzhubei I, Jordan DM, Sunyaev SR (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* Chapter 7: t7-t20.
16. Kondrashov FA (2005) Prediction of pathogenic mutations in mitochondrially encoded human tRNAs. *Hum Mol Genet* 14(16): 2415-2419.
17. Ruiz-Pesini E, *et al.* (2007) An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Res* 35(Database issue): D823-D828.
18. He Y, *et al.* (2010) Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature* 464(7288): 610-614.
19. Li M, *et al.* (2010) Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. *Am J Hum Genet* 87(2): 237-249.
20. Goto H, *et al.* (2011) Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. *Genome Biol* 12(6): R59.