Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals

Kaixiong Ye,1 Jian Lu,2 Fei Ma,3 Alon Keinan4, and Zhenglong Gu1,5

A majority of mitochondrial DNA (mtDNA) mutations reported to be implicated in diseases are heteroplasmic, a status with coexisting mtDNA variants in a single cell. Quantifying the prevalence of mitochondrial heteroplasmy and its pathogenic effect in healthy individuals could further our understanding of its possible roles in various diseases. A total of 1,085 human individuals from 14 global populations have been sequenced by the 1000 Genomes Project to a mean coverage of ~2,000× on mtDNA and a combination of stringent thresholds and a maximum-likelihood method to define heteroplasmy, we demonstrated that ~90% of the individuals carry at least one heteroplasmy. At least 20% of individuals harbor heteroplasmies reported to be implicated in disease. Mitochondrial heteroplasmy tend to show high pathogenicity, and is significantly overrepresented in disease-associated loci. Consistent with their deleterious effect, heteroplasmies with derived allele frequency larger than 60% within an individual show a significant reduction in pathogenicity, indicating the action of purifying selection. Purifying selection on heteroplasmies can also be inferred from nonsynonymous and synonymous heteroplasm comparison and the unfolded site frequency spectra for different functional sites in mtDNA. Nevertheless, in comparison with population polymorphic mtDNA mutations, the purifying selection is much less efficient in removing heteroplasmic mutations. The prevalence of mitochondrial heteroplasmy with high pathogenic potential in healthy individuals, along with the possibility of these mutations drifting to high frequency inside a subpopulation of cells across lifespan, emphasizes the importance of managing mitochondrial heteroplasmy to prevent disease progression.

Significance

There are hundreds to thousands of copies of mitochondrial DNA (mtDNA) in each human cell in contrast to only two copies of nuclear DNAs. These mtDNAs can differ from each other as a result of inherited or somatic mutations. The coexistence of multiple mtDNA variants in a single cell or among cells within an individual is called heteroplasmy. Mitochondrial heteroplasmy has been shown to be implicated in a large spectrum of human diseases. Besides classical mitochondrial diseases such as mitochondrial myopathy, myoclonic epilepsy with ragged red fibers, and mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes, mitochondrial heteroplasmy also plays roles in complex disorders, including myopathy, lactic acidosis, and stroke-like episodes, mitochondrial encephalomyopathy with ragged red fibers, and mitochondrial encephalomyopathies such as mitochondrial myopathy, myoclonic epilepsy, and autism. Mitochondrial heteroplasmy has been shown to be implicated in a large spectrum of human diseases. Nevertheless, in comparison with population polymorphic mtDNA mutations, the purifying selection is much less efficient in removing heteroplasmic mutations. The prevalence of mitochondrial heteroplasmy with high pathogenic potential in healthy individuals, along with the possibility of these mutations drifting to high frequency inside a subpopulation of cells across lifespan, emphasizes the importance of managing mitochondrial heteroplasmy to prevent disease progression.

Author contributions: K.Y., A.K., and Z.G. designed research; K.Y., J.L., and Z.G. performed research; K.Y. contributed new reagents/analytic tools; K.Y. and F.M. analyzed data; and K.Y., J.L., F.M., A.K., and Z.G. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

1To whom correspondence may be addressed. Email: zg27@cornell.edu or ky279@cornell.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1403521111/-/DCSupplemental.
characterizing the pathogenic potential of mitochondrial heteroplasmy in healthy individuals and its underlying evolutionary forces will further our understanding of the roles of mtDNA variations in aging, tumorigenic, and neurodegenerative processes. In this study, we addressed this issue by analyzing deep sequencing data of mtDNA for 1,085 healthy individuals sampled from 14 global populations in the 1000 Human Genomes Project (25). First, we quantified the prevalence of mitochondrial heteroplasmy, especially disease-associated heteroplasmy, in this healthy cohort. We further characterized the pathogenicity of mitochondrial heteroplasmy with computationally predicted and experimentally reported pathogenic effects. Moreover, we scrutinized the patterns of genomic distribution and site-frequency spectrum for mitochondrial heteroplasmy and elucidated the major evolutionary forces underlying these patterns. We demonstrated that pathogenic mitochondrial heteroplasmy is prevalent in healthy individuals, likely due to insufficient purifying selection in removing them. The implication of our results in health management was also discussed.

Results
Mitochondrial Heteroplasmy Is Prevalent in the Normal Human Population. The average depth of coverage for 1,085 individuals sequenced by ILLUMINA or SOLID in the 1000 Genomes Project is 1,805× (SI Appendix, Fig. S1), allowing the identification of low-frequency heteroplasms. We applied a combination of stringent thresholds to define heteroplasmy with high confidence and estimated the frequency of heteroplasmy with a multinomial Genetin likelihood (ML) method. In total, we identified 4,342 heteroplasms. There were nine individuals included in our analysis that were additionally sequenced by LS454, providing an opportunity of verifying the accuracy of our computational pipeline. For the 22 heteroplasms identified in these nine individuals with the ILLUMINA data, all of them were observed in the LS454 data with similar frequency (SI Appendix, Fig. S2), reassuring the reliability of our computational procedure. With the 4,342 heteroplasms identified in 1,085 individuals, 973 individuals (89.68%) have at least one heteroplasmy. In an extreme case, an individual (HG00740) carried 71 heteroplasies (Fig. 1A). The population prevalence of heteroplasmy depends on the criteria of defining heteroplasmy. The higher the cutoff for MAF, the lower the prevalence. However, even with MAF cutoffs of 5% and 10%, heteroplasmy is observed in 63.50% and 44.42% of the individuals, respectively (SI Appendix, Fig. S3).

The majority of heteroplasies are present at low frequency (Fig. 1B). The median ML-estimated MAF is 2.71%. The skew to low frequency is similar to the site frequency spectrum of population polymorphism but less severe (SI Appendix, Fig. S4). These heteroplasms were observed at 2,531 mtDNA sites across different regions in mtDNA, and 1,757 (69.42%) of these sites are heteroplasmic in only one individual (Fig. 1C and SI Appendix, Fig. S4B). Among all heteroplasmic sites, 36.67% were also polymorphic in the population (permutation test, P < 1.00e-5; SI Appendix, Fig. S4 C and D). There is a positive correlation between the population incidence of heteroplasmy and the population MAF of polymorphic sites (linear regression R² = 0.2558, P < 2.20e-16; SI Appendix, Fig. S5). In a previous study, a relative mutation rate for each site in the mitochondrial genome was defined as the absolute frequency of mutation occurrence in a phylogenetic tree constructed with global human samples (26). Using this dataset, we showed that heteroplasmic sites have significantly higher relative mutation rates than homoplasmic sites (Wilcoxon rank-sum test, P < 2.20e-16; SI Appendix, Fig. S6A), and that the incidence of heteroplasmy is positively correlated with relative mutation rate (linear regression R² = 0.3702, P < 2.20e-16; SI Appendix, Fig. S6B). These observations further confirmed the reliability of our pipeline in identifying heteroplasms and indicated that high mutation rate might be a major driving force for the population prevalence of heteroplasmy.

Mitochondrial Heteroplasmy Is Overrepresented in Disease-Associated Sites. Of the 4,342 detected heteroplasms, 301 (7.11%) are reported to be disease-associated (7) and 210 individuals (19.35% of 1,085) carried at least one disease-associated heteroplasmy. These observations prompted us to further investigate the disease implication for these heteroplasms. Among the 13,639 mtDNA sites that satisfied quality control criteria and were examined in our study, 399 (2.93%) are disease associated (7). However, the corresponding number is 147 (5.81%) among the 2,531 heteroplasmic sites, which is significantly higher than expected by chance ($\chi^2$ test, P = 2.52e-12). The percentage of disease-associated sites among population polymorphic sites (6.30%) is also significantly more than random expectation ($\chi^2$ test, P = 1.44e-14) but is comparable to that of heteroplasmic sites (Fig. 1C). For the two disease categories that have the highest number of associated sites, mitochondrial myopathy and mitochondrial encephalomyopathy, heteroplasmy is overrepresented, even compared with polymorphism. Among all of the sites examined, 64 (0.47%) have been reported to be associated with mitochondrial myopathy, and 52 (0.38%) with mitochondrial encephalomyopathy. Only one site is shared...
Mitochondrial heteroplasmy is highly pathogenic. To explore the pathogenicity of mitochondrial heteroplasmic variants, we applied various methods to predict the deleterious effect of nonsynonymous (NS) and tRNA mutations. For NS heteroplasmies, we first defined pathogenicity scores for all possible NS changes in the mtDNA with the MutPred algorithm (27, 28). The pathogenicity score ranges from 0 to 1 with a higher pathogenicity score indicating greater likelihood of being pathogenic. For all possible 24,206 NS changes in the mtDNA, the average pathogenicity score is 0.63 (SD = 0.15). For all 1,184 NS heteroplasmies in the dataset, the average score is 0.63 (SD = 0.16), similar to random NS mutations. In contrast, the average pathogenicity score for all 467 population polymorphisms is 0.52 (SD = 0.16), significantly lower than that of heteroplasmies and all possible variants (P = 4.24e-36 and 5.96e-55, respectively; Fig. 2B). To quantify the pathogenic potential of heteroplasmic variants in comparison with population polymorphic ones, we could choose a cutoff of pathogenicity score and define NS variants with scores higher than this cutoff as pathogenic. To avoid the arbitrary use of cutoffs, we applied a series of cutoffs from 0.6 to 0.8 and, in general, heteroplasmies are 1.87–4.26x more likely to be pathogenic than polymorphism (SI Appendix, Fig. S7). As a verification, the pathogenicity of NS variants was also predicted using PolyPhen-2 (29, 30). PolyPhen-2 yielded comparable predictions with MutPred (SI Appendix, Fig. S8). The percentage of damaging heteroplasmic variants is significantly lower than random expectation, but significantly higher than that of polymorphic ones (Fig. 2C).

We further investigated the pathogenicity of heteroplasmies in tRNA genes. We used the pathogenic prediction for all possible variants in tRNA genes from a previous study which used evolutionary information in functional assessment (31). The percentage of pathogenic variants for heteroplasmies is 64.23%, significantly lower than that for all possible variants (75.58%, χ² test, P = 0.0023) but significantly higher than that for polymorphism (33.33%, χ² test, P = 2.63e-06). In other words, tRNA heteroplasmies are 1.93 (95% CI: 1.51–2.63) times more likely to be pathogenic than polymorphisms. Similar trends were observed when we separated tRNAs into three regions: loop, Watson–Crick pairing positions in stem, and non-Watson–Crick pairing positions in stem (Fig. 2D).

Mitochondrial Heteroplasmy Is Subject to Purifying Selection. The high pathogenicity of heteroplasmies and their strong association with diseases suggest that heteroplasmies might be under purifying selection. To test this hypothesis, we investigated the genome-wide distribution of heteroplasmies, their unfolded site frequency spectra and the relationship between pathogenicity and derived allele frequency (DAF).

We first examined synonymous and NS variants in protein-coding genes: 5.78% of all possible NS changes in mtDNA were observed with heteroplasmies, which is significantly lower than that of synonymous changes (8.10%, χ² test, P = 2.10e-20; Fig. 3A and SI Appendix, Figs. S4 and S5). Similarly, for mitochondrial encephalomyopathy, heteroplasmies in the control region is comparable to that for synonymous heteroplasmies. In comparison with these two types of sites, the distributions of DAF for NS, tRNA, and rRNA heteroplasmies are significantly shifted toward lower frequencies (Wilcoxon rank-sum test, P < 9.42e-16; Fig. 3B). Furthermore, heteroplasmies within the tRNA stem tend to have a lower DAF frequency than those in the loop and heteroplasmies in the tRNA loop and anticondon loop regions have significantly lower DAF than those in other tRNA regions (SI Appendix, Fig. S9). Intriguingly, heteroplasmies at disease-associated sites also exhibit significantly lower DAF than that of synonymous heteroplasmies (Wilcoxon rank-sum test, P = 2.07e-10; Fig. 3C). Taken together, these results suggest purifying selection is acting on functional heteroplasmies to keep them at low frequency.

The effect of purifying selection on removing deleterious heteroplasmies suggests a possible reverse correlation between the level of pathogenicity and the frequency of a heteroplasm. Consistent with this expectation, as depicted in Fig. 3D, heteroplasmies with low derived frequency inside an individual tend to have high pathogenicity scores. This negative relationship can be modeled with a logistic function ($R^2 = 0.9794$, $P < 9.76e-06$). From the regression, we inferred that the pathogenicity scores are comparable among heteroplasmies with DAF less than 60% and declines as DAF exceeds 60%. This pattern indicates that pathogenic heteroplasmies must reach high frequency before they are selected against, and 60% in general might be a good estimate of the threshold for pathogenic heteroplasmic mutations to express deleterious effect. Consistent with the impact of purifying selection on removing deleterious heteroplasmies, our results also show that though heteroplasmies observed in a few individuals (1–4) have comparable pathogenicity scores (mean = 0.64, SD = 0.16), those observed in more than five individuals have significantly lower pathogenicity scores.
The selection function for heteroplasmy also follows an exponential decay ($R^2 = 0.9650, P = 4.65e-06$; Fig. 4B). Interestingly, in contrast to polymorphism, it has an additional constant very close to 1, indicating that purifying selection is too weak to effectively remove pathogenic heteroplasmas, likely due to their low frequency inside the cells. Using the selection functions for both polymorphisms and heteroplasmas, the relative effect of selection on two different amino acid variations could be assessed by a ratio of the exponential functions for the two pathogenicity scores (28). For example, a population polymorphic variant with a pathogenicity score of 0.8 is subject to $\sim 2x$ stronger purifying selection than a polymorphic variant with a score of 0.6. In comparison, the strength of purifying selection on two heteroplasmas with pathogenicity scores of 0.8 and 0.6 is almost the same. This quantitative comparison further confirms that purifying selection on heteroplasmy is much less efficient than that on population polymorphism in removing deleterious mutations.

**Discussion**

Next-generation sequencing technologies enable the detection of mitochondrial heteroplasmy at the genome-wide level with unprecedented resolution. However, specificity of detection and accuracy of quantification can only be achieved when sequencing errors and technical artifacts are carefully controlled for. A set of criteria for detecting heteroplasmy with modern sequencing technologies have been developed in a few pioneering studies (13, 16, 21, 32). Integrating criteria that have been proven to be effective (SI Appendix, Table S2), our computational pipeline filtered low-quality bases and unreliable mappings, especially minimizing the complications of nuclear mitochondrial sequen-
ces (NumtS) (33); it also used double-stranded validation, which required heteroplasmy to be detected in both strands with support from multiple reads. Furthermore, our computational pipeline estimated the frequency of heteroplasmy with a maximum likelihood method by taking into account sequencing error and yielded a log likelihood ratio (LLR) indicating the confidence of true positive heteroplasmy. The applications of these tested criteria ensure the correct detection and accurate quantification of heteroplasmy. The reliability of our computational pipeline was confirmed by examining nine individuals sequenced by both ILLUMINA and LS454 (SI Appendix, Fig. S2 and Dataset S1). Moreover, the biologically meaningful patterns of mitochondrial heteroplasmy observed in our study also augment the reliability of our computational pipeline. The complete list of

**Purifying Selection Is Less Efficient on Heteroplasmy Than on Polymorphism.** Although heteroplasmy sites show evidence of purifying selection, we hypothesized that purifying selection on heteroplasmy is much weaker than that on polymorphism due to the low frequencies of most heteroplasmas inside individual cells. Indeed, consistent with this hypothesis, the difference between the percentages of synonymous and NS variants is much bigger for polymorphisms than heteroplasmas ($\chi^2$ test, $P < 2.2e-16$; Fig. 4A). To further quantitatively compare the effect of natural selection on these two types of mtDNA variants, we defined a selection function by dividing the observed distribution of pathogenicity scores for all NS heteroplasmas (or polymorphisms) by the expected distribution of pathogenicity scores from all possible NS variants. This quantitative method has been previously applied to mitochondrial polymorphisms (28). In the absence of natural selection, mutations are similar to random draws from all possible changes in the genome, so the selection function is expected to be equal to a constant, 1. Consistent with previous study (28), the selection function of polymorphism can be modeled by a simple function of exponential decay ($R^2 = 0.9758, P = 4.71e-06$; Fig. 4B). The parameterizations in our data are similar to the previous study (28). We also confirmed that the observed value for polymorphisms with very high pathogenicity scores (>0.9) deviates from the exponential fit, indicating that forces other than purifying selection might have acted on these variants (28).

### Fig. 3.

- Purifying selection on mitochondrial heteroplasmy. (A) The prevalence of synonymous and NS heteroplasmas, which is defined as the percentage of all possible (synonymous or NS) changes that is observed to be heteroplasmic. **$P = 2.01e-10$ in $\chi^2$ test. (B) The distribution of DAF for heteroplasmas in different mtDNA genomic regions. (C) The distribution of DAF for disease-associated and synonymous heteroplasmas. (D) The average pathogenicity score in each bin of DAF. Error bar represents 1 SE. The red line represents model-fitting with a logistic function of $y = 0.67/(1 + e^{-1.0-4.18})$. $R^2 = 0.9794, P = 9.76e-06.$

(mean = 0.43, SD = 0.25, Wilcoxon rank-sum test, $P = 8.74e-04$; SI Appendix, Fig. S10). When we examined the DAF of pathogenic tRNA heteroplasmas, we also found that 81.90% of heteroplasmas with DAF less than 5% are pathogenic, whereas only 30% of heteroplasmas with DAF larger than 95% are pathogenic (Fisher’s exact test, $P = 0.0010$).

### Fig. 4.

- Less-efficient purifying selection on mitochondrial heteroplasmy than on polymorphism. (A) The prevalence of synonymous and NS heteroplasmas in comparison with that of synonymous and NS polymorphisms. (B) The selection function for heteroplasmy (polymorphism) defined by dividing the observed distribution of pathogenicity scores for heteroplasmy by the expected distribution of pathogenicity scores from all possible NS variants. The dashed line represents the expected value, 1, for selection function under neutral evolution. The exponential fit for polymorphism is $y = 12e^{-\frac{x}{0.21}}, R^2 = 0.9758, P = 4.71e-06.$ The exponential function for heteroplasmy is $y = 10e^{-\frac{x}{0.672}} + 0.99, R^2 = 0.9650, P = 4.65e-06.$
heteroplasmies identified in our study can be found in Dataset S2. Additionally, we did not observe consistent and significant population or sex difference in heteroplasm patterns (SI Appendix, Figs. S11 and S12).

The prevalence of mitochondrial heteroplasmy at genomewide scale has been explored in a few studies with smaller sample size and shallower sequencing depth. From the 1000 Genomes Pilot Project, 163 individuals were sequenced to 37.7–3,535× coverage and 45% were observed to possess heteroplasmic sites with MAF mostly larger than 10% (12). Another study sequenced 114 individuals with ILLUMINA to a mean coverage of 67× and 17 individuals to a mean coverage of 211×. Among these 131 individuals, 24.43% were detected to possess heteroplasmy with MAF larger than 10% (13). Moreover, a study used a 454 Genome Sequencer FLX system and sequenced 40 HapMap individuals to a mean coverage of 120×, and 65% individuals were found to have heteroplasmies with MAF higher than 9% (14). With a MAF cutoff of 10%, the prevalence of heteroplasmy is 44.42% in our dataset (SI Appendix, Fig. S3), which is within the range of previous estimates and very close to the estimate from the 1000 Genomes Pilot Project (12). Our study benefits from higher coverage and is able to detect heteroplasmy with MAF as low as 1%. With a much larger sample size, we estimate that the prevalence of heteroplasmy in the healthy population is at least 90%. Because the majority of heteroplasmies is present at very low frequency (Fig. 1B), it is very likely that heteroplasy is universal to all healthy individuals. Results from a recent study conducted on a small sample support this idea (17).

The high pathogenic potential of mitochondrial heteroplasmy is consistently demonstrated with experimentally observed disease-associated mutations, computationally predicted functional effect, and the presence of weak negative selection. First, experimentally reported diseases-associated mtDNA mutations are overrepresented in both polymorphic and heteroplasmic sites (Fig. S4). This pattern has been previously observed in a study with a much smaller sample size (13), and it suggests that heteroplasmic and polymorphic variants are either only mildly deleterious or not yet effectively removed by purifying selection. Because polymorphic variants have gone through generations of purifying selection, their overrepresentation in disease-associated sites is likely resulted from their mild deleterious effect. In contrast, because heteroplasmic variants have a much shorter time frame for natural selection, they are likely subject to weaker purifying selection and have higher pathogenic potential. However, the overrepresentation of or polymorphism and heteroplasmy in disease-associated sites may also reflect the research bias toward using known polymorphic sites in disease studies. Second, we artificially created all possible variants in the mitochondrial genome and computationally predicted their pathogenic effects, which serve as a pathogenicity benchmark before being subject to purifying selection. In comparison with this theoretical expectation, heteroplasmy has slightly lower pathogenicity, whereas polymorphism has much lower effect (Fig. 2 B–D), which is consistent with the fact that polymorphism has been subject to generations of purifying selection and only variants with mild deleterious effect could survive; it also suggests that though purifying selection also acts on heteroplasmy, its strength may be weak and therefore the pathogenic effect of heteroplasmy is very close to the theoretical expectation without purifying selection. Last, we observed convincing signals of purifying selection on heteroplasmy and demonstrated its weaker strength than that on polymorphism, further supporting the high pathogenic potential of heteroplasmy.

The prevalence of pathogenic heteroplasmic mtDNA mutations in the general population due to inefficient purifying selection has important clinical implication. Although only ~1 in 5,000 people suffers from mitochondrial diseases (24), the incidence of pathogenic mtDNA mutations could be much higher because of the mitochondrial threshold effect that masks the deleterious effect of low-frequency pathogenic mutations. A study of 10 common pathogenic mtDNA mutations revealed an incidence of at least 1 in 200 subjects (23). For these 10 mutations, the prevalence of heteroplasmy in our samples is 1 in 155 (95% CI: 83–356). When we included all identified disease-associated mtDNA mutations (7), the incidence of pathogenic heteroplasmy is 19.35%; or 1 in 5 individuals (95% CI: 4.62–5.87). Given the likely underestimation of disease–mtDNA mutation association and the observed prevalence of heteroplasmic mtDNA mutations with high predicted pathogenic scores in this study, the real frequency of pathogenic mitochondrial heteroplasmy could be much higher than this estimation.

Multiple underlying mechanisms have been proposed to modulate the expansion of deleterious mtDNA mutations at the cellular level. According to computational modeling of the relaxed replication of mtDNA in both dividing and nondividing cells, even with random generational growth rate typical of most observers, an individual is more than enough for low-frequency heteroplasmy to reach high frequency or even homoplasy in a small population of cells (34–36). On average it only takes ~70 generations of cell divisions to reach homoplasy from a new mutation; that is only ~25 y for epithelial cells, which experience three cell turnovers per year (34). In postmitotic tissues, such as skeletal muscle and neurons, the mean time to homoplasy is ~40 y (35, 36). Besides random genetic drift during intracellular mitochondrial turnover and cell divisions (34, 35, 37), natural selection with replicative or survival advantage has been proposed to either accelerate or slow down the spread of pathogenic mutations (38–40). Extensive experimental observations have recorded abundant clonally expanded mtDNA mutations in human tissues, especially in aged individuals (40, 41). More importantly, both computational modeling and experimental evidence support that mutation accumulated with age results mostly from the clonal expansion of mutations that existed early in life, rather than de novo mutations later in life (35, 37, 41, 42). All individuals included in the 1000 Genomes Project were healthy at the time of sample collection (25). The prevalence of pathogenic mitochondrial heteroplasmy in healthy individuals observed in this study raises the concern that they could expand to high frequency in a fraction of cells later in life, exceed the critical phenotypic threshold, and lead to age-related diseases. Future studies are needed to unravel the mechanisms of clonal expansion of pathogenic heteroplasmy, to elucidate the roles of mitochondrial heteroplasmy in complex disorders, and to develop effective strategies in managing these mutations to prevent the progression into disease.

Materials and Methods

Sequencing Data. Sequencing reads mapped to the mitochondrial genome in the 1000 Genomes Project phase 1 data were downloaded from the 1000 Genomes Project data server. Our analysis focused on 1,085 unrelated individuals from 14 populations, which were sequenced using either ILLUMINA or SOLID platforms. There were nine individuals sequenced by two methods (ILLUMINA and LS454). These individuals were used to confirm the reliability of our computational pipeline with ILLUMINA data. See SI Appendix, Table S1 for more detailed information.

Computational Pipeline for Calling Heteroplasmy and Polymorphism. An expanded version of the methods is in SI Appendix. Briefly, sequencing reads retrieved from the 1000 Genomes Project data server were remapped to the human genome, both nuclear and mitochondrial genomes, using GSNAP (43). Only reads uniquely mapped to the mitochondrial genome were recorded to minimize the complications of NumtS (33). We further filtered the data and defined “usable sites” based on the following three quality control criteria: (i) Phred quality score ≥20 for used bases; (ii) 10× coverage of qualified bases on both positive and negative strands; (iii) 95% individuals satisfy criteria i and ii. Together, 13,639 mtDNA sites satisfied these quality control criteria and were examined in our study. A candidate heteroplasmic site was defined with the following two criteria: (i) the raw frequency for the major allele is no less than 1% or dominant strand; and (ii) all alleles have support from at least two reads on each strand.

For each candidate heteroplasmic site, we further applied a ML method to accurately estimate the frequency of the major allele while taking into account sequencing error (32, 44). For example, for all bases mapped to the positive
strand of a site, i bases are the major alleles and k bases are the minor alleles. Each base has respective sequencing quality, corresponding to the probability as assembled using the alleles present at homoplasmic sites, and the major identify polymorphic sites. A consensus sequence for each individual was assembled for each individual and compared among all individuals to on both strands. After detecting heteroplasmy, consensus sequences were of 4,342 heteroplasmies were defined; among them, 153 have a minor allele We estimated Ye et al. PNAS

\[ L(f) = \prod_{j=1}^{n} \left(1 - f_j \right) \left(1 - \epsilon_j \right) + f_j \epsilon_j. \]

We estimated \( f \) under two models: heteroplasm (m,) and homoplasy (m), and a LRR was calculated as \( \log(L(m)/L(m_0)) \). A high-confidence heteroplasm was defined as candidate heteroplasm with LRR no less than 5 (32). With all these criteria (see SI Appendix, Table S2 for a brief list), a total of 4,342 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 less than 1% on both strands. After detecting heteroplasm, consensus sequences were assembled for each individual and compared among all individuals to identify polymorphic sites. A consensus sequence for each individual 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 population.

The Measure of Pathogenicity. The pathogenicity scores for all possible NS changes, inferred based on the revised Cambridge Reference Sequence, were predicted with the MutPred algorithm (27); as a verification, their pathogenenic effects were further predicted with PolyPhen-2 (28). Both methods yielded comparable results. The MutPred pathogenicity scores were retrieved from a previous study (28). A higher pathogenicity score indicates a higher likelihood that the NS change is pathogenic (27, 28). The pathogenic effect of tRNA mutations were obtained from a previous publication (31). Disease association information was obtained from MITOMAP (7).

ACKNOWLEDGMENTS. We thank Mr. Paul Billing-Ross and Drs. Andrew Clark, Jason Locasale, Patrick Stover, and Lin Xu for their discussions and comments on the manuscript. This work was supported by various funds from Cornell University, an International Life Sciences Institute future leader award, National Science Foundation Grant MCB-1243588, and National Institutes of Health Grant 1R01A1085286 (to Z.G.). K.Y. is a Center for Vertebrate Genomics Scholar at Cornell University.
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 server 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 heteroplasy 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 \( \varepsilon \). The underlying parameter of interest is
the frequency of the major allele \( f \). The likelihood function could be written as follows:

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

We estimated \( f \) under two models: heteroplasmy (\( m_1 \)) and homoplasmy (\( m_0 \)). And a log-
likelihood ratio (LLR) was calculated as \( \log \left( \frac{L(f_{m_1})}{L(f_{m_0})} \right) \). 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 heteroplasy, 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 heteroplasy, 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 heteroplasy
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 heteroplasms 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.

![Graph showing the positive correlation between the incidence of heteroplasmy and the minor allele frequency of polymorphism in the sample of 1085 individuals.](image)

**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 represents 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.
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 heteroplasms. 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>
<thead>
<tr>
<th>Population</th>
<th># by ILLUMINA</th>
<th># by SOLID</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASW</td>
<td>50</td>
<td>11</td>
<td>61</td>
</tr>
<tr>
<td>CEU</td>
<td>81</td>
<td>0</td>
<td>81</td>
</tr>
<tr>
<td>CHB</td>
<td>81</td>
<td>16</td>
<td>97</td>
</tr>
<tr>
<td>CHS</td>
<td>92</td>
<td>8</td>
<td>100</td>
</tr>
<tr>
<td>CLM</td>
<td>50</td>
<td>10</td>
<td>60</td>
</tr>
<tr>
<td>FIN</td>
<td>75</td>
<td>18</td>
<td>93</td>
</tr>
<tr>
<td>GBR</td>
<td>70</td>
<td>19</td>
<td>89</td>
</tr>
<tr>
<td>IBS</td>
<td>6</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>JPT</td>
<td>78</td>
<td>11</td>
<td>89</td>
</tr>
<tr>
<td>LWK</td>
<td>82</td>
<td>14</td>
<td>96</td>
</tr>
<tr>
<td>MXL</td>
<td>52</td>
<td>12</td>
<td>64</td>
</tr>
<tr>
<td>PUR</td>
<td>52</td>
<td>3</td>
<td>55</td>
</tr>
<tr>
<td>TSI</td>
<td>98</td>
<td>0</td>
<td>98</td>
</tr>
<tr>
<td>YRI</td>
<td>76</td>
<td>12</td>
<td>88</td>
</tr>
</tbody>
</table>

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, Nigeria. Populations highlighted in blue are those of European ancestry.
Table S2. Comparison of criteria for calling heteroplasmy.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequencing Technology</td>
<td>ILLUMINA ~16,700X</td>
<td>ILLUMINA 36 &amp; 76 bp ~67 &amp; ~211X</td>
<td>ILLUMINA ~1,170X</td>
<td>ILLUMINA, Agilent, NimbleGen</td>
<td>ILLUMINA, SOLID</td>
</tr>
<tr>
<td>Mapping Tools</td>
<td>Eland</td>
<td>MIA</td>
<td>BWA</td>
<td>GSNAP</td>
<td>GSNAP remapping</td>
</tr>
<tr>
<td>Mismatches</td>
<td>&lt;= 3 in 36 bp</td>
<td>Default</td>
<td>Default</td>
<td>Default</td>
<td>Default or 7%</td>
</tr>
<tr>
<td>Reads</td>
<td>remove low-quality reads</td>
<td>remove duplicate reads &amp; low-quality reads</td>
<td>--</td>
<td>--</td>
<td>Only reads originally mapped to mtDNA by 1000 genome project</td>
</tr>
<tr>
<td>Mapping</td>
<td>--</td>
<td>--</td>
<td>Unique</td>
<td>Unique</td>
<td>Unique</td>
</tr>
<tr>
<td>Base Quality</td>
<td>&gt;=23 for all bases in the read</td>
<td>&gt;= 20 on site; &gt;= 15 for 5 bp flanking</td>
<td>&gt;= 30 on site</td>
<td>&gt;=20 on site</td>
<td>&gt;=20 on site</td>
</tr>
<tr>
<td>Minimum Depth</td>
<td>&gt;= 10 distinct reads</td>
<td>--</td>
<td>&gt;= 100X HQ depth on each strand</td>
<td>&gt;=20 X</td>
<td>&gt;=10X HQ depth on each strand</td>
</tr>
<tr>
<td>Double Strand Validation (control for context-dependent error &amp; PCR duplicate during sequencing)</td>
<td>&gt;=3 reads on each strand</td>
<td>&gt;= 1 read on each strand</td>
<td>&gt;=100 HQ depth on each strand; &gt;= 2% raw frequency on each strand;</td>
<td>--</td>
<td>&gt;= 2 reads on each strand; &gt;=1% raw frequency on each strand</td>
</tr>
<tr>
<td>Minor allele frequency</td>
<td>&gt;=1.6%</td>
<td>&gt;=10%</td>
<td>&gt;= 2% on each strand</td>
<td>--</td>
<td>&gt;=1% on each strand</td>
</tr>
<tr>
<td>Log-likelihood ratio</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>&gt;=5</td>
<td>&gt;=5</td>
</tr>
</tbody>
</table>

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: