Title: Germline selection shapes human mitochondrial DNA diversity*

Authors: Wei Wei[1,2], Salih Tuna[3,4], Michael J Keogh[1], Katherine R Smith[5†], Timothy J Aitman[6,7], Phil L Beales[8,9], David L Bennett[10], Daniel P Gale[11], Maria A K Bitner-Glindzicz[8,9,12,#], Graeme C Black[13,14], Paul Brennan[15,16,17], Perry Elliott[18,19], Frances A Flinter[20,21], R Andres Floto[22,23,24], Henry Houlden[25], Melita Irving[21], Ania Koziell[26,27], Eamonn R Maher[28,29], Hugh S Markus[30], Nicholas W Morrell[4,22], William G Newman[13,14], Irene Roberts[31,32,33], John A Sayer[16,34], Kenneth G C Smith[22], Jenny C Taylor[33,35], Hugh Watkins[35,36,37], Andrew R Webster[38,39], Andrew O M Wilkie[33,37,40], Catherine Williamson[41,42], on behalf of the NIHR BioResource - Rare Diseases[+] and the 100,000 Genomes Project - Rare Diseases Pilot[+], Sofie Ashford[4,43], Christopher J Penkett[3,4], Kathleen E Stirrups[3,4], Augusto Rendon[3,5†], Willem H Ouwehand[3,4,44,45,46¶], John R Bradley[4,22,24,29,47¶], F Lucy Raymond[4,28¶], Mark Caulfield[5,48†], Ernest Turro[3,4,49§], Patrick F Chinnery[1,2,4§]

Affiliations:

[1] Department of Clinical Neurosciences, School of Clinical Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK.

[2] Medical Research Council Mitochondrial Biology Unit, Cambridge Biomedical Campus, Cambridge, UK.

[3] Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK.

[4] NIHR BioResource, Cambridge University Hospitals NHS Foundation, Cambridge Biomedical Campus, Cambridge, UK.

[5] Genomics England, Charterhouse Square, London, UK.

[6] MRC Clinical Sciences Centre, Faculty of Medicine, Imperial College London, London, UK.

[7] Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK.

[8] Genetics and Genomic Medicine Programme, UCL Great Ormond Street Institute of Child Health, London, UK.

[9] Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK.

[10] The Nuffield Department of Clinical Neurosciences, University of Oxford, John Radcliffe Hospital, Oxford, UK.

[11] UCL Centre for Nephrology, University College London, London, UK.

[12] University College London, London, UK.

[13] Evolution and Genomic Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK.

[14] Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester Universities Foundation NHS Trust, Manchester, UK.

[15] Newcastle University, Newcastle upon Tyne, UK.

[16]Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK.

[17]Northern Genetics Service, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK.

[18]UCL Institute of Cardiovascular Science, University College London, London, UK.

[19]Barts Heart Centre, St Bartholomew's Hospital, Barts Health NHS Trust, London, UK.

[20]Guy's and St Thomas' Hospital, Guy's and St Thomas' NHS Foundation Trust, London, UK.

[21]Clinical Genetics Department, Guy's and St Thomas NHS Foundation Trust, London, UK.

[22]Department of Medicine, School of Clinical Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK.

[23]Royal Papworth Hospital NHS Foundation Trust, Cambridge, UK.

[24]Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK.

[25]Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK.

[26]King's College London, London, UK.

[27]Department of Paediatric Nephrology, Evelina London Children's Hospital, Guy's & St Thomas' NHS Foundation Trust, London, UK.

[28]Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK.

[29]NIHR Cambridge Biomedical Research Centre, Cambridge Biomedical Campus, Cambridge, UK.

[30]Stroke Research Group, Department of Clinical Neurosciences, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK.

[31]MRC Molecular Haematology Unit, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK.

[32]Department of Paediatrics, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK.

[33]NIHR Oxford Biomedical Research Centre, Oxford University Hospitals Trust, Oxford, UK.

[34]Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK.

[35]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK.

[36]Department of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK.

[37]Oxford University Hospitals NHS Foundation Trust, Oxford, UK.

[38]Moorfields Eye Hospital NHS Foundation Trust, London, UK.

[39]UCL Institute of Ophthalmology, University College London, London, UK.

[40]MRC Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK.

[41]Division of Women's Health, King's College London, London, UK.

[42]Institute of Reproductive and Developmental Biology, Surgery and Cancer, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK.

[43]Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK.

[44]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK.

[45]NHS Blood and Transplant, Cambridge Biomedical Campus, Cambridge, UK.

[46]British Heart Foundation Cambridge Centre of Excellence, University of Cambridge, Cambridge, UK.

[47]Department of Renal Medicine, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK.

[48]William Harvey Research Institute, NIHR Biomedical Research Centre at Barts, Queen Mary University of London, London, UK.

[49]MRC Biostatistics Unit, Cambridge Institute of Public Health, University of Cambridge, Cambridge, UK.

[§]Corresponding author. Email: pfc25@cam.ac.uk & et341@cam.ac.uk

[+] a full list of the authors is included in the Supplementary Material under ‘Collaborators’

¶ = equal contribution

† = equal contribution

# = deceased

§ = corresponding authors

* This manuscript has been accepted for publication in Science. This version has not undergone final editing. Please refer to the complete version of record at http://www.sciencemag.org/. The manuscript may not be reproduced or used in any manner that does not fall within the fair use provisions of the Copyright Act without the prior, written permission of AAAS.

Structured Abstract

INTRODUCTION

Only 2.4% of the 16.5Kb mitochondrial genome (mtDNA) shows homoplasmic variation at >1% frequency in humans. Migration patterns have contributed to geographic differences in the frequency of common genetic variants, but population genetic evidence implicates selection shaping the evolving mtDNA phylogeny. It is not clear how and when this occurs.

Unlike the nuclear genome, mtDNA is maternally transmitted and there are many copies in each cell. Initially a new genetic variant only affects a proportion of the mtDNA (heteroplasmy). A reduction in the amount of mtDNA per cell during female germ cell development causes a genetic bottleneck. This leads to the rapid segregation of mtDNA molecules and different heteroplasmy levels between siblings. Although primarily governed by random genetic drift, there is evidence of selection occurring during this process in animals. However, it has been difficult to demonstrate this convincingly in humans.

RATIONALE

To determine whether there is selection for or against heteroplasmic mtDNA variants during transmission we studied 12,975 whole genome sequences, including 1,526 mother-offspring pairs where 45.1% had heteroplasmy affecting >1% of mtDNA molecules. Harnessing both the mtDNA and nuclear genome sequences, we then determined whether the nuclear genetic background influenced mtDNA heteroplasmy, validating our findings in another 40,325 individuals.

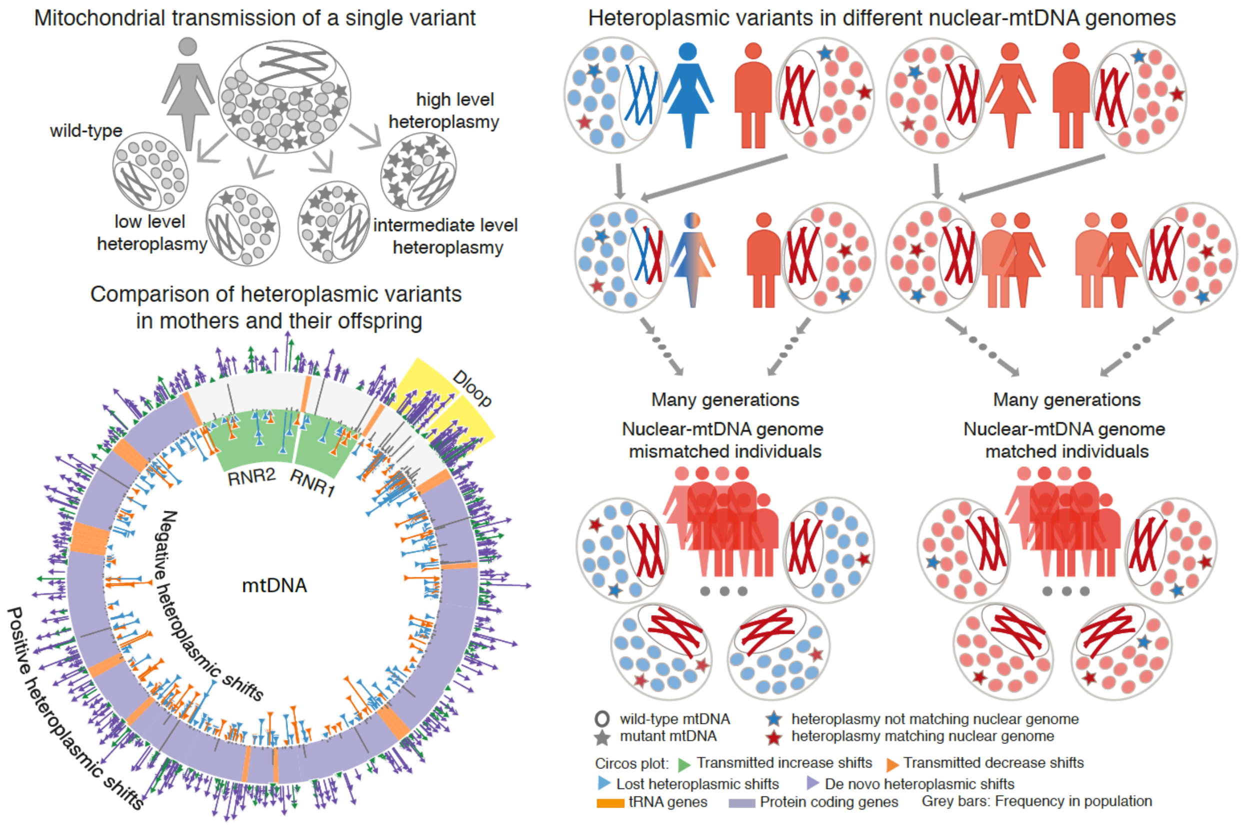
RESULTS

Novel mtDNA variants were less likely to be inherited than known variants, where the level of heteroplasmy tended to increase on transmission. Variants in the ribosomal RNA genes were less likely to be transmitted, and variants in the non-coding displacement (D)-loop were more likely to be transmitted. MtDNA variants predicted to affect the protein sequence tended to have lower heteroplasmy levels than synonymous variants. In 12,975 individuals, there was a correlation between the location of heteroplasmic sites and known D-loop polymorphisms, including the absence of variants in critical sites required for mtDNA transcription and replication.

We defined 206 unrelated individuals where the nuclear and mitochondrial genomes were from different human populations. In these, new population-specific heteroplasmies were more likely to match the nuclear genetic ancestry than the mitochondrial genome on which the mutations occurred. These findings were independently replicated in 654 informative individuals.

CONCLUSION

The characteristics of mtDNA in the human population are shaped by selective forces acting on heteroplasmy within the female germ line, and are influenced by the nuclear genetic background. The signature of selection can be seen over one generation, ensuring consistency between these two independent genetic systems.


Mitochondrial transmission of a single variant
wild-type
high level heteroplasmy
low level heteroplasmy
intermediate level heteroplasmy
Comparison of heteroplasmic variants in mothers and their offspring
Dloop
RNR2
RNR1
mtDNA
Negative heteroplasmic shifts
Positive heteroplasmic shifts
Heteroplasmic variants in different nuclear-mtDNA genomes
Many generations
Nuclear-mtDNA genome mismatched individuals
Many generations
Nuclear-mtDNA genome matched individuals
wild-type mtDNA
mutant mtDNA
heteroplasmy not matching nuclear genome
heteroplasmy matching nuclear genome
Circos plot: Transmitted increase shifts
Transmitted decrease shifts
Lost heteroplasmic shifts
De novo heteroplasmic shifts
tRNA genes
Protein coding genes
Grey bars: Frequency in population

Germline selection of human mitochondrial DNA is shaped by the nuclear genome. Top left: Initially new mtDNA variants are heteroplasmic, with the proportion changing during maternal transmission. Bottom left: Selection during the transmission of mtDNA heteroplasmy. Transmitted variants are more likely have been seen before as homoplasmic polymorphisms. Right: New haplogroup-specific variants are more likely to match the nuclear genetic ancestry than the mtDNA ancestry.

Abstract: ~2.4% of the human mitochondrial genome (mtDNA) shows common homoplasmic genetic variation. Analyzing 12,975 whole genome sequences we show that 45.1% of individuals from 1,526 mother-offspring pairs harbor a mixed population of mtDNA (heteroplasmy), but the propensity for maternal transmission differs across the mitochondrial genome. Over one generation, we observe selection both for and against variants in specific genomic regions, and previously seen variants were more likely to be transmitted. New heteroplasmies were more likely to match the nuclear genetic ancestry than the mitochondrial genome on which the mutations occurred, validating our findings in 40,325 individuals. Thus, human mtDNA at the population level is shaped by selective forces within the female germline under nuclear genetic control to ensure consistency between the two independent genetic lineages.

One Sentence Summary: Human mitochondrial DNA (mtDNA) undergoes selection in the female germ line which is shaped by the nuclear genome.

Primarily inherited from the maternal line, the 16.5Kb human mitochondrial DNA (mtDNA) genome acquired mutations sequentially following the emergence of modern humans out of Africa (*1-3*). Pedigree and phylogenetic analyses have estimated a *de novo* mtDNA nucleotide substitution rate of ~10^{-8}/base pair/year (*4*). However, from 30,506 mitochondrial genome sequences from across the globe (*5*), only 2.4% of nucleotides show genetic variation with frequencies greater than 1% within a population (**Fig. 1**). Although contentious (*6, 7*), selection could explain the non-random distribution of common variants across the mitochondrial genome in the human population.

Heteroplasmic mtDNA variants are common and maternally inherited

We analyzed high-depth mtDNA sequences from 1,526 mother-offspring pairs (mean depth in the mothers = 1,880x, range 249x-7,454x; mean depth in the offspring = 1,901x, range 259x-7,475x; mothers *vs*. offspring, P=0.49, two-sample t-test) **(fig. S1)**. We called homoplasmic and heteroplasmic mtDNA variants from whole-blood DNA sequence data (*8, 9*) and filtered out heteroplasmic calls likely to be due to errors (*9*, *10, 11*). We identified a mixed population of mtDNA (heteroplasmic variants) with a heteroplasmic variant allele frequency (VAF) >1% with high confidence in 47.8% of mothers (1,043 heteroplasmic variants at 812 sites) and 42.5% of offspring (893 heteroplasmic variants at 693 sites) (**Fig. 1**, **table S1 and Data S1**). In 22 individuals, where the whole genome was independently sequenced twice, the heteroplasmic mtDNA calls were 96.4% concordant (**fig. S2**) (*9*). As expected (*12, 13*), there was a small but significant positive correlation between the number of heteroplasmic variants and age of mother (P=6.42 x 10^{-11}, R^2=0.17, CI= 0.12 - 0.23, Pearson's correlation) (**fig. S3**), with mothers having more heteroplasmic variants than offspring (mean number in the mothers = 0.68, range 0-6;

mean number in the offspring = 0.58, range 0-4; *P*=0.002, effect size = 0.68, Wilcoxon rank sum test) (**Fig. 2A**).

We defined three categories of heteroplasmic variants: (1) transmitted/inherited, if the variant was present in the mother and the offspring and was heteroplasmic in at least one of the two; (2) lost, if the heteroplasmic variant was present in the mother but not detectable in the offspring; and (3) *de novo*, if the heteroplasmic variant was present in the offspring but not detectable in the mother (**table S1***)* (*9*). Note that very low level heteroplasmies (<1% VAF) may be missed by our sequencing and bioinformatics pipeline. Hence, "lost" and "*de novo*" variants could potentially be present at very low levels in, respectively, the offspring's and mother's germline. The heteroplasmic fraction (HF) of transmitted heteroplasmic variants (mean HF = 19.5%, sd = 13.9%) was significantly higher than the HF of lost variants (mean HF = 5.6%, sd = 6.3%) in the mothers (*P*<2.2 x 10^{-16}, effect size = 4.24, Wilcoxon rank sum test); and the HF of inherited heteroplasmic variants (mean HF = 19.8%, sd = 14.1%) was significantly higher than the HF of *de novo* variants (mean HF = 6.2%, sd = 7.4%) in the offspring (*P*<2.2 x 10^{-16}, effect size = 4.06, Wilcoxon rank sum test) (**Fig. 2B and table S1**). The HF of transmitted variants in the offspring strongly correlated with the corresponding maternal level (*P*=1.52 x 10^{-93}, R^2=0.79, CI=0.75 - 0.82, Pearson's correlation) (**Fig. 2C**). In total, 477 *de novo* heteroplasmic variants were observed at >1% HF in the offspring that were not seen in the mother, in keeping with previous estimates (*13*). To ensure these data were not due to technical errors, we determined whether any heteroplasmic variants in the offspring were also present in their fathers. Amongst 313 father-offspring pairs, the offspring harbored 196 heteroplasmic variants with HF >1%, and only one of these was also observed in the corresponding father. This was a common population variant (population minor allele frequency (MAF) = 25.8% (*5*)) in the D-loop region (m.152T>C) which was homoplasmic in the father and had an HF of 12.4% in his child. The alternate allele was not detected in the mother, suggesting this is a recurrent site of mutation or conceivably due to the paternal transmission of mtDNA.

The difference between HF in mothers and their offspring can be measured in percentage points (**Fig. 2D**)(*14*). This metric is limited by the difference between the HF of the mother and the boundaries 0 and 100% and the magnitude of the percentage change does not correspond with the magnitude of the fold-change in VAF. For example, a change from 50% to 55% would be given the same value as a change from 1% to 6%, even though the latter implies 6-fold increase in the proportion of mtDNA carrying the alternate allele. We therefore studied the log2 ratio of HF between offspring and mothers after imputation of HF values below 1% to our detection threshold of 1% (subsequently termed the heteroplasmy shift (HS), **Fig. 2, E and F**) (*9*), which shrunk HSs towards zero only when the true HF in either the mother or the offspring was below 1%.

Overall, there was no significant difference between the number of heteroplasmic variants with a positive (n=731) and a negative (n=798) HS (*P*=0.091, binomial test). The HS distribution around zero was moderately symmetric and gave a marginal *P* value for asymmetry (*P*=0.05, one sample *t*-test) (**Fig. 2, D and E**), consistent with random segregation of mitochondria during meiosis (*14, 15*). All of the HSs were <6 in magnitude, corresponding to a <64-fold increase or decrease in HF across one generation, with 3 exceptions. *De novo* variants at m.57T>C (HF=99.3%), m.8993T>G (HF=82.1%) and m.14459G>A (HF=93.6%) were detected in three unrelated offspring and not present in the corresponding mothers (**figs. S4 to S6**). m.14459G>A is a non-synonymous (NS) variant in *ND6* which, on the basis of evidence from previously published pedigrees (*16, 17*), causes Leber hereditary optic neuropathy (LHON) and Leigh syndrome/dystonia. m.8993T>G is a NS variant in *ATP6* (L156R), which has been observed on many independent occasions in Leigh syndrome or neurogenic ataxia with retinitis pigmentosa (*18-20*). Although these extreme HSs could reflect differences in the mechanism of transmission for pathogenic mtDNA mutations (*21*), ascertainment is a more likely explanation because childhood-onset neurodegenerative diseases were recruited as part of this study (*22*). Ascertainment bias is unlikely to explain the *de novo* occurrence of m.57T>C, but these findings indicate that extreme HSs at moderate HFs are not typical of human populations.

As expected, the non-coding displacement (D)-loop had the highest substitution frequency (7.64×10^{-5} /base/genome/transmission) of all the regions in the mitochondrial genome (**Fig. 3A and table S2**)(*13*). In total, we observed 16 out of 57 previously defined (*5*) pathogenic mutations in the 1,526 mother-offspring pairs (**Fig. 3B**). After excluding m.14459G>A and m.8993T>G, where the extreme HS likely reflects ascertainment bias, the mean HS for the remaining 14 pathogenic mutations was not significantly different from zero (P=0.22, one sample t-test), nor from the mean HS for the remaining 1,076 non-pathogenic variants (P=0.11, two sample t-test). Thus, overall we did not see a strong signature of selection for or against pathogenic alleles, although our statistical analysis does not preclude that a subset of the observed pathogenic alleles may be under selection. Intriguingly, only three mothers carried the most common heteroplasmic pathogenic mutation m.3243A>G (*23*), each with a low HF (5.2%, 3.6% and 1.7%), which decreased in the corresponding offspring, to levels falling below our detection threshold in two of the three offspring (3.9%, <1% and <1%). Six of the 16 pathogenic mutations were not detectable in the mothers, giving a *de novo* mutation rate for known pathogenic mutations of 393/100,000 live births (95% CI 144 – 854), which is ~3.7-fold higher than previously reported (*24*).

To gain insight into possible mutational mechanisms, we determined the trinucleotide mutational signature. As shown previously, C>T and T>C substitutions were the most common type of substitution in homoplasmic variants (*5*) and cancer somatic mtDNA mutations (*25*). For heteroplasmic variants, C>T and T>C substitutions were also predominant, although we also observed a small but significant excess of C>A, C>G, T>A and T>G substitutions ($P<2.2 \times 10^{-16}$, odd ratio = 0.36, CI = 0.29 - 0.44, Fisher's exact test) (**fig. S7)**. Given that the heteroplasmic variant signature was not identical to the homoplasmic variant signature, this suggests that the germline transmission shapes the mutational signatures seen in homoplasmic variants at the population level. Also of note, *de novo* mutations were more likely to involve a CpG-containing trinucleotide ($P=3.01 \times 10^{-6}$, odd ratio = 0.50, CI = 0.38 - 0.66, Fisher's exact test) (**Fig. 3C**). Although controversial (*26*), this could be because methylation of NpCpG sites on the mtDNA genome predisposes to *de novo* mtDNA mutations, as seen in the nuclear genome.

Known mtDNA variants are more likely to be transmitted than novel

We then compared heteroplasmic variants which have been seen before in the general population (known) and those not previously observed (novel). Variants were considered novel if they were absent from the 1000 Genomes datasets and dbSNP and were seen in at most one individual amongst 30,506 NCBI mtDNA sequences (*5*). Novel heteroplasmic variants were 4.7-fold less commonly transmitted from mother to offspring than known variants ($P=3.55 \times 10^{-13}$, odd ratio=2.60, CI=1.97 - 3.45, Fisher's exact test), and the HS for transmitted known variants was more likely to be positive ($P=0.0002$, probability=0.40, CI=0.35 - 0.45, binomial test) (**Fig. 3, D and E**). Also, the transmitted heteroplasmic variants were more likely to affect known haplogroup-specific sites (*27*) compared to the lost and *de novo* heteroplasmic variants ($P=7.86 \times 10^{-11}$, odds ratio=0.40, CI=0.30-0.53, and $P=0.0016$, odds ratio=0.62, CI=0.46-0.84, respectively, Fisher's exact test) (**Fig. 3F**). This suggests that factors may modulate the transmission of mtDNA heteroplasmy within the female germline over a single generation and influence the likelihood that they become established within human mtDNA populations. As heteroplasmic variants are acquired throughout life, they must be removed at transmission to offspring at a higher rate than they appear *de novo* as, otherwise, each generation would be accompanied by an expected increase in the number of heteroplasmic variants which may be deleterious (*28*). In keeping with this, the number of novel variants present in the mother but not transmitted (lost variants), exceeded the number of *de novo* novel variants detected in the offspring (P=7.93 x 10^{-7}, probability=0.62, CI= 0.57 - 0.67, binomial test) (**Fig. 3D**), in part reflecting the accumulation of heteroplasmic variants with increasing age in the mothers (**fig. S3**).

Selection for and against heteroplasmy in different genomic regions

We analyzed different functional regions of the genome and found evidence indicating region-specific selection for or against heteroplasmic variants. The distributions of HF in the 1,526 mother-offspring pairs were significantly different between the D-loop, rRNA, tRNA, and coding regions (**Fig. 4A and table S3**). Within the coding region, the NS and synonymous (SS) variants also had different distributions (P=2.74 x 10^{-5}, Kolmogorov-Smirnoff test). The NS/SS ratio was greater for the heteroplasmic variants than for the homoplasmic variants (P=3.98 x 10^{-24}, odds ratio=1.91, CI=1.68 - 2.18, Fisher's exact test), and the *de novo* and lost heteroplasmic variants had a higher NS/SS than the transmitted variants (transmitted *vs de novo*: P=0.0056, odds ratio=1.69, CI=1.15 - 2.48; transmitted *vs* lost: P=0.01, odds ratio=1.57, CI=1.10 - 2.24, Fisher's exact test) (**Fig. 4B**). The heteroplasmic variants were more often in conserved sites than the homoplasmic variants (P=3.71 x 10^{-77}, odds ratio=3.21, CI=2.86 - 3.60, Fisher's exact test), and the transmitted heteroplasmic variants were less conserved than the *de novo* (P=0.0018, odds ratio=1.62, CI=1.19 - 2.22, Fisher's exact test) and lost (P=9.60 x 10^{-9}, odds ratio=2.25, CI=1.69 - 3.03, Fisher's exact test) heteroplasmic variants **(Fig. 4C)**. Also, heteroplasmic variants with a positive HS were less conserved than those with a negative HS (P=0.03, odds ratio=1.28, CI=1.01 - 1.61, Fisher's exact test). Variants in the rRNA genes were more likely to show a decrease in the heteroplasmy level on transmission than an increase (P=1.00 x 10^{-4}, probability=0.65, CI=0.57 - 0.72, binomial test) (**Fig. 4D**), and the mean HS was significantly less than zero (P=8.21 x 10^{-5}, d=0.30, one sample *t*-test) (**Fig. 4E**).

In order to understand the determinants of transmission of heteroplasmic variants with a reduced risk of confounding, we used multi-variable logistic regression to model the probability of transmission across all 1,526 mother-offspring pairs (*9*). We modeled the transmission probability of a variant as a function of its HF in the mother, the identity of the mitochondrial genome region containing it, and its known *vs* novel status (**Fig. 4, F to H and Fig. 3D**) (*9*). The probability that a heteroplasmic variant in the mother was transmitted to her offspring was associated with its HF in the mother (P<2.2 x 10^{-16}, coefficient estimate=1.17, sd=0.08, logistic regression) (**Fig. 4F**). Variants in the D-loop were more likely to be transmitted (P=0.04, coefficient estimate=0.39, sd=0.19, logistic regression) than average and those in the rRNA were less likely to be transmitted (P=0.0026, coefficient estimate=-0.94, sd=0.31, logistic regression) than average (**Fig. 4G**). The novel variants were less likely to be transmitted than the known variants (P=0.028, coefficient estimate=0.43, sd=0.19, logistic regression) (**Fig. 3D**), even after accounting for all other covariates, including HF in the mothers.

Heteroplasmic variants in the non-coding Displacement (D-) loop

To cast light on the possible effects of selection on the non-coding D-loop, we derived a high-resolution map of heteroplasmic variants in 12,975 individuals, which included the 1,526 mother-offspring pairs (mean mtDNA genome depth = 1832x, sd=945x; mean depth of D-loop = 1569x, sd=819x) (**Fig. 5, A to C and fig. S8) (***9*). We found an association between the homoplasmic allele frequency amongst 30,506 NCBI mtDNA sequences and the proportion of individuals heteroplasmic for the same allele ($P<2.2 \times 10^{-16}$, logistic regression) (**Fig. 5, A to C**) similar to that previously observed (*5*). Of the 17 regions in the D-loop (**Fig. 5C** bottom - purple and orange bars), two had a significantly greater number of heteroplasmic variants than expected by chance. These regions correspond to the proposed replication fork barrier associated with the D-loop termination sequence (MT-TAS2) (*29*) and MT-CSB1 (MT-TAS2: $P=4.5 \times 10^{-11}$, odds ratio=0.40, CI=0.30 - 0.54; MT-CSB1: $P=7.0 \times 10^{-6}$, odds ratio=0.39, CI=0.24 - 0.61, Fisher's exact test *vs* remainder of the D-loop).

To help understand the evolution of the D-loop, we identified all the heteroplasmic variants not identified on mtDNA phylogenies across a subset of 10,210 unrelated individuals from the original dataset (*9*). Five of these heteroplasmic variants were shared by more than one individual and were present exclusively in people with a particular haplogroup (**Fig. 5, D and E**). One variant (m.16237A>T) was present in multiple individuals from two different branches of the phylogeny (L0a1&2 and M35b2) (**Fig. 5, D and E**). Compared to homoplasmic sequences from across the world (*5*), only m.299C>A was observed previously as a homoplasmic variant (in 3/30,506 individuals), each time on the R30b1 haplogroup background. This suggests that individuals we saw who were heteroplasmic for m.299C>A (**Fig. 5F**), also descended from the same maternal ancestor as the three homoplasmic individuals seen previously (*5*), but belonged to a closely related maternal lineage that had not yet reached fixation. These recurrent heteroplasmies contributed to the distinct trinucleotide mutational signature of the D-loop ($P=2.3 \times 10^{-137}$, Stouffer's method for combining Fisher *P* values), which involves prominent noncanonical substitutions, and is consistent with the conclusion that the homoplasmic trinucleotide mutational signature of mtDNA is shaped by germline transmission of heteroplasmic variants **(Fig. 5D and fig. S9**).

We observed an absence of low-level heteroplasmic variants in critical sites required for the initiation of mtDNA transcription and replication. These zones include several conserved sequence boxes and the light strand promoter (MT-LSP: *P*=7.7 x 10^{-18}, odds ratio=10.12, CI=5.43 – 20.31, Fisher's exact test), which are required for mtDNA transcription and mtDNA replication (*30*). Certain regions with no known function (*31*) (eg. 16,400-16,500; **Fig. 5C**) also had a complete lack of low-level heteroplasmic variants, which suggests that an intact sequence at these regions is essential for mitochondrial function, perhaps genome propagation. The coordinates of the conserved and non-conserved regions provide a guide for functional studies of the mtDNA D-loop which has been incompletely characterized to date.

The nuclear genetic background influences the heteroplasmy landscape

Most of the ~1,500 known mitochondrial proteins are synthesized from the nuclear genome, including the majority of polypeptide subunits of the oxidative phosphorylation system, and the machinery required to replicate and transcribe the mitochondrial genome *in situ* (*1*). Selection for or against specific mtDNA variants must therefore occur in the context of a specific nuclear genetic background. To explore this, we identified 12,933 individuals for whom a confident mtDNA haplogroup could be predicted (**fig. S10**). We compared the haplogroup of each individual with the corresponding nuclear genetic ancestry, and identified three distinct groups of individuals: (1) a haplogroup matched group (n=11,867, 91.7%) where the mtDNA haplogroup was concordant with the nuclear ancestry; (2) a mismatched group (n=295, 2.3%) where the nuclear ancestry and mtDNA were from different human populations; and, (3) a group where the nuclear ancestry could not be reliably determined (n=771, 6.0%) (**Fig. 6, A and B and fig. S10**). Subsequent analyses focused on the haplogroup matched and mismatched groups (*9*).

8,159 heteroplasmic variants at 3,854 of the 16,569 distinct sites on the mitochondrial genome were present in the matched group, and 195 heteroplasmic variants at 163 distinct sites were present in the mismatched group. The mean number of heteroplasmic variants and mean HF were not statistically different between the matched and mismatched groups (**fig. S11**). Next, we studied distinct heteroplasmic sites in the 10,179 of 12,933 individuals who were not related on the basis of their nuclear genome (9,414 in the matched group, 217 in the mismatched group and 548 in the other group). Distinct heteroplasmic sites were more likely to affect known haplogroup specific sites (*27*) than the rest of the mitochondrial genome (P<2.2 x 10^{-16}, Fisher's exact test), particularly within the mismatched group (P=0.001, odds ratio=1.70, CI=1.22 - 2.36, Fisher's exact test) (**Fig. 6C**).

We extracted 2,641 haplogroup-specific variants present in only one super-population (European, Asian or African) on the world mtDNA phylogeny (*27*). We built a predictive model of transmission of these variants using logistic regression in 9,385 unrelated European and Asian nuclear ancestries using 2,215 European (n=940) and Asian (n=1,275) specific variants on the mtDNA phylogeny, omitting the Africans because of the diversity and small number (**figs. S10 and S12**)(*9*). We included the super-population and the logit population allele frequency as covariates. We also included a dummy variable indicating whether or not the variant matched the mitochondrial ancestry of the individual carrying the variant. Finally, for the matched and mismatched groups, we included a separate variable indicating whether or not the variant super-population matched the nuclear ancestry of the individual who carried the variant.

We fitted the model to 768 heteroplasmic variants in 9,179 unrelated matched individuals and 30 heteroplasmic variants in 206 unrelated mismatched individuals (*9*). The heteroplasmic variants in the mismatched group were significantly more likely to match the ancestry of the nuclear genetic background than the mtDNA background on which the heteroplasmy occurred (*P*=2.9 x 10^{-4}, coefficient estimate=0.85, sd=0.24, logistic regression, **Fig. 6D and table S4**). These findings suggest that the new mtDNA variants underwent selection to match the nuclear genome. Given the high mutation rate of the mitochondrial genome and the patterns we observed over one generation, the selective process is likely to occur within the female germline.

To independently validate this finding, we repeated this analysis with an additional 40,325 WGS recruited through the Genomics England 100,000 Genomes Rare Disease Main Programme (*9*). There were 36,038 individuals in a haplogroup matched group, 1,098 in a haplogroup mismatched group, and 3,124 in a group where the nuclear ancestry could not be reliably determined (**figs. S12, S13**). As before, we focused on the European and Asian specific variants observed in 23,931 unrelated European and Asian individuals. We fitted the same logistic regression model to 1,942 heteroplasmic variants in 23,277 unrelated matched individuals, and 67 heteroplasmic variants in 654 unrelated individuals where the nuclear and mtDNA had a different ancestral origin. Again, the heteroplasmic variants in the mismatched group were more likely to match the ancestry of the nuclear genetic background than the ancestral background of the mtDNA on which the heteroplasmy occurred (*P*=1.33 x 10^{-3}, coefficient estimate=0.47, se=0.15, logistic regression, **Fig. 6D and table S4**). An inverse-weighted meta-analysis of the discovery and validation cohorts yielded a significant association across the two datasets (*P*=3.3 x $10^{-6,}$ coefficient estimate=0.59, se=0.13). To gain a better understanding of underlying mechanisms we studied the gene location and HF of 97 heteroplasmic variants identified in the mismatched groups across both the discovery and validation studies. Potentially functional variants were found in the non-coding region and RNA genes, and also included 14 non-synonymous protein coding variants in the *MT-ATP, MT-COX, MT-CYB* and *MT-ND* regions (**fig. S14**). This raises the possibility that differences in oxidative phosphorylation and ATP synthesis are responsible the association we observed.

Discussion

Several explanations have been proposed for the high substitution rate of the non-coding mtDNA D-loop, including a high intrinsic mutation rate, and/or a permissive sequence relative to the coding regions (*31*). Here we show that the segregation of mtDNA heteroplasmy likely plays a role in shaping D-loop population polymorphisms by a mechanism operating within the female germline. Similar findings have been seen in *Drosophila* where D-loop variants 'selfishly' drive segregation favoring a specific mtDNA genotype (*32*). These observations have implications for the development of mitochondrial transfer techniques for preventing the inheritance of severe pathogenic mtDNA mutations in humans (*33, 34*). After mitochondrial transfer, ~15% of human embryonic stem cell lines show reversion to the original mtDNA genotype (*34-36*). The reasons for this are not fully understood, but the selective propagation of D-loop heteroplasmy is a plausible explanation. Our findings implicate the nuclear genome in this process. This places greater emphasis on matching both nuclear and mtDNA backgrounds when selecting potential mitochondrial donors, in order to minimize the possibility of nuclear-mitochondrial incompatibility following mitochondrial transfer.

In cases of heteroplasmic mtDNA, one allele can be preferentially copied, or segregate to high levels in a population of daughter cells. This can lead to changes in mtDNA allele frequency during the lifetime of an individual cell, tissue or organism through genetic drift (*38,* *39*). A high mtDNA content buffers fluctuations in allele frequency. However, if the number of copies falls to a low level, this creates a ‘genetic bottleneck’, increasing the possibility of large changes in allele frequency.

There is a ~1000-fold reduction in cellular mtDNA content during human germ cell development (*40*) is followed by a period of intense proliferation and migration when the germ cells migrate to form the developing gonad (*41*). This process is dependent on oxidative phosphorylation, and is accompanied by a massive increase in mtDNA levels (*40*). Under these conditions, variants that compromise mitochondrial ATP synthesis will be selected against. On the other hand, variants that promote mtDNA replication will have an advantage, potentially explaining the preferential transmission of specific D-loop variants. Subtle selective pressures will have maximal impact at this time, so the nuclear genetic influence we observed will most likely come into to play during this critical period of development. Our analysis was based on whole-blood DNA, and not a direct analysis of the germ line, so it is possible that tissue-specific differences in heteroplasmy come in to play. However, by studying 53,300 mtDNA sequences including mother-child pairs, our findings indicate that human mtDNA at the population level is influenced by selective forces acting within the female germline and modulated by the nuclear genetic background. These are apparent within one generation, and ensure consistency between these two independent genetic systems, shaping the current world mtDNA phylogeny.

Materials and Methods:

Participants, approvals and sequence acquisition

The primary data was whole genome sequencing (WGS) from whole-blood DNA from 13,037 individuals in the NIHR BioResource - Rare Diseases and 100,000 Genomes Project Pilot studies (**table S5)** (*22*) After quality control (QC, see below and (*9*)), 12,975 samples including 1,526 mother-offspring pairs were included in this study. For demographics see (*9*). Ethical approval was provided by the East of England Cambridge South national research ethics committee (REC) under reference number: 13/EE/0325. WGS was performed using the Illumina TruSeq DNA PCR-Free sample preparation kit (Illumina, Inc.) and an Illumina HiSeq 2500 sequencer, generating a mean depth of 45x (range from 34x to 72x) and greater than 15x for at least 95% of the reference human genome (**fig. S8A**).

Extracting mitochondrial sequences, quality control and variant detection

WGS reads were aligned to the Genome Reference Consortium human genome build 37 (GRCh37) using Isaac Genome Alignment Software (version 01.14; Illumina, Inc.). Reads aligning to the mitochondrial genome were extracted from each BAM file and analyzed using MToolBox (v1.0) (*8*, *9*). Variant Call Files and the merged VCF were normalized with bcftools and vt (*44, 45, 46)*, and duplicated variants were dropped with vt. The final VCF was annotated using the Variant Effect Predictor (VEP) (*47*). Further QC was carried out as described (*9*). Potential DNA cross-contamination was investigated using verifyBamID (*48*) in the nuclear genome, and mtDNA variant calls (*9*).

Determining matched and mismatched groups

The pairwise relatedness and nuclear ancestry were estimated using nuclear genetic markers as described (*9*). MtDNA haplogroup assignment was performed using HaploGrep2 (*27, 57*). We then compared the mtDNA phylogenetic haplogroup with the nuclear genetic ancestry in the same individual, and identified three distinct groups of individuals as described in the text.

Defining novel variants

Variants were considered to be novel if absent from 1000 Genomes datasets and dbSNP and were seen in at most one individual amongst 30,506 NCBI mtDNA sequences (*5*).

MtDNA mutational spectra and signature

Mutational spectra were derived from the reference and alternative alleles as described (*25, 58*).

Probability of maternal mtDNA transmission

We modelled the probability of transmission of heteroplasmic variants observed in the mothers using the following logistic regression model:

$$\text{logit}\, P\left(y_{ijl} = 1\right) = \alpha + \beta_1 \mathbf{1}_{j=1} + \beta_2 \mathbf{1}_{j=2} + \beta_3 \mathbf{1}_{j=3} + \beta_4 \mathbf{1}_{j=4} + \gamma w_{ijl} + \eta z_{ijl}$$

where $y_{ijl} = 1$ if the *l*th variant within mitochondrial genomic region j in mother *i* was transmitted and zero otherwise; $j = 0, 1, 2, 3$ or 4 denote the coding, Dloop, rRNA, tRNA and the remainder sequences, respectively; w_{ijl} is the logit of the HF of the *l*th variant within mitochondrial genomic region j in mother *i*; and $z_{ijl} = 1$ if the *l*th variant within mitochondrial genomic region j in mother *i* was observed in no individuals from the 1000 Genomes datasets, dbSNP and at most one individual amongst 30,506 NCBI mtDNAs, otherwise it was equal to zero.

Homoplasmic allele frequency in the population and heteroplasmic variants

We fitted a logistic regression model to explore the relationship between the homoplasmic allele frequency in the general population and the rate at which individuals who are not homoplasmic for the alternate allele are heteroplasmic (*9*).

Defining haplogroup specific variants on mtDNA phylogenetic tree

We extracted 4,476 SNVs present on mtDNA phylogenetic tree (*27*), then focused on SNVs either present in only one super-population (European, Asian or African), or present on two super-populations but commonly seen in one population (>1%) and not seen or extremely rare in the other population in 17,520 mtDNAs (*5*). This defined 2,641 haplogroup-specific variants, including 426 African variants, 1,275 Asian variants and 940 European variants.

Nuclear genome ancestry and mtDNA heteroplasmic variants

We modelled the presence or absence of a heteroplasmic variant in a particular individual using logistic regression. We considered only the 2,215 mtDNA variants from of over 4,000 haplogroup-specific variants that are present exclusively in European or Asian branches of the world mtDNA phylogeny (*27*), as this allows unambiguous assignation of mitochondrial ancestry to each variant. To avoid the potential for bias induced by recent shared ancestry between individuals, we considered only the 9,631 unrelated individuals in matched and mismatched groups. We fitted the following logistic regression model:

$$\text{logit}\, P\left(y_{ij} = 1\right) = \alpha + \beta_1 \mathbf{1}_{\boldsymbol{x_j=1}} + \beta_2 \mathbf{1}_{\boldsymbol{x_j=2}} + \gamma w_j + \eta \mathbf{1}_{\boldsymbol{z_i=w_i}} + \omega \mathbf{1}_{\boldsymbol{x_j=w_i \cap z_i=w_i}} + \psi \mathbf{1}_{\boldsymbol{x_j=w_i \cap z_i \neq w_i}}$$

where $y_{ij} = 1$ if variant j is heteroplasmic in individual i, and zero otherwise; $x_j = 0, 1$ or 2 depending on whether the variant ancestry is Asian, African or European, respectively; w_j is the logit of the homoplasmic allele frequency of variant j in 30,506 NCBI samples; $z_i = 0, 1$ or 2 depending on whether the mitochondrial ancestry of individual *i* is Asian, African or European, respectively; and $w_i = 0, 1$ or 2 depending on whether the nuclear ancestry of individual *i* is Asian, African or European, respectively. The indicator variable **1** evaluates to 1 if the conditions in its subscript are met and zero otherwise.

Validation dataset

We repeated the nuclear-mtDNA ancestry analysis in 42,799 WGS from whole-blood DNA in the Genomics England 100,000 Genomes Rare Disease Main Programme aligned to GRCh37 or/and hg38 using the same bioinformatics pipeline. See (*9*) for details.

References and Notes:

1. S. B. Vafai, V. K. Mootha, Mitochondrial disorders as windows into an ancient organelle. *Nature* **491**, 374-383 (2012).
2. D. C. Wallace, Mitochondrial DNA variation in human radiation and disease. *Cell* **163**, 33-38 (2015).
3. J. B. Stewart, P. F. Chinnery, The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nature Reviews Genetics* **16**, 530-542 (2015).
4. P. Soares *et al.*, Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet* **84**, 740-759 (2009).
5. W. Wei, A. Gomez-Duran, G. Hudson, P. F. Chinnery, Background sequence characteristics influence the occurrence and severity of disease-causing mtDNA mutations *PLoS Genet* **14**, e1007364 (2018).
6. B. Cavadas *et al.*, Fine Time Scaling of Purifying Selection on Human Nonsynonymous mtDNA Mutations Based on the Worldwide Population Tree and Mother-Child Pairs. *Hum Mutat* **36**, 1100-1111 (2015).
7. E. Ruiz-Pesini, D. Mishmar, M. Brandon, V. Procaccio, D. C. Wallace, Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science* **303**, 223-226 (2004).
8. C. Calabrese *et al.*, MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. *Bioinformatics* **30**, 3115-3117 (2014).
9. See Supplementary Methods on line
10. M. Gerstung, E. Papaemmanuil, P. J. Campbell, Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics* **30**, 1198-1204 (2014).
11. M. Gerstung *et al.*, Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat Commun* **3**, 811 (2012).
12. M. Wachsmuth, A. Hubner, M. Li, B. Madea, M. Stoneking, Age-Related and Heteroplasmy-Related Variation in Human mtDNA Copy Number. *PLoS Genet* **12**, e1005939 (2016).

13. B. Rebolledo-Jaramillo *et al.*, Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA. *Proc Natl Acad Sci U S A* **111**, 15474-15479 (2014).

14. M. Li *et al.*, Transmission of human mtDNA heteroplasmy in the Genome of the Netherlands families: support for a variable-size bottleneck. *Genome Res* **26**, 417-426 (2016).

15. P. F. Chinnery *et al.*, The inheritance of mitochondrial DNA heteroplasmy: random drift, selection or both? *Trends Genet* **16**, 500-505 (2000).

16. D. M. Kirby, S. G. Kahler, M. L. Freckmann, D. Reddihough, D. R. Thorburn, Leigh disease caused by the mitochondrial DNA G14459A mutation in unrelated families. *Ann Neurol* **48**, 102-104 (2000).

17. J. M. Shoffner *et al.*, Leber's hereditary optic neuropathy plus dystonia is caused by a mitochondrial DNA point mutation. *Ann Neurol* **38**, 163-169 (1995).

18. I. J. Holt, A. E. Harding, R. K. Petty, J. A. Morgan-Hughes, A new mitochondrial disease associated with mitochondrial DNA heteroplasmy. *Am J Hum Genet* **46**, 428-433 (1990).

19. Y. Tatuch *et al.*, Heteroplasmic mtDNA mutation (T----G) at 8993 can cause Leigh disease when the percentage of abnormal mtDNA is high. *Am J Hum Genet* **50**, 852-858 (1992).

20. J. M. Shoffner *et al.*, Lebers Hereditary Optic Neuropathy Plus Dystonia Is Caused by a Mitochondrial-DNA Point Mutation. *Annals of Neurology* **38**, 163-169 (1995).

21. I. J. Wilson *et al.*, Mitochondrial DNA sequence characteristics modulate the size of the genetic bottleneck. *Hum Mol Genet* **25**, 1031-1041 (2016).

22. The NIHR BioResource, on behalf of the 100,000 Genomes Project, Whole-genome sequencing of rare disease patients in a national healthcare system. bioRxiv doi: 10.1101/507244

23. G. S. Gorman *et al.*, Prevalence of nuclear and mtDNA mutations related to adult mitochondrial disease. *Ann Neurol* **77**, 753-759 (2015).

24. H. R. Elliott, D. C. Samuels, J. A. Eden, C. L. Relton, P. F. Chinnery, Pathogenic mitochondrial DNA mutations are common in the general population. *Am J Hum Genet* **83**, 254-260 (2008).

25. Y. S. Ju *et al.*, Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *eLife* **3**, (2014).

26. Y. Matsuda, I. Hanasaki, R. Iwao, H. Yamaguchi, T. Niimi, Estimation of diffusive states from single-particle trajectory in heterogeneous medium using machine-learning methods. *Phys Chem Chem Phys* **20**, 24099-24108 (2018).

27. M. van Oven, M. Kayser, Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat* **30**, E386-394 (2009).

28. H. J. Muller, The relation of recombination to mutational advance. *Mutat Res* **1**, 2-9 (1964).

29. Y. Shi *et al.*, Mitochondrial transcription termination factor 1 directs polar replication fork pausing. *Nucleic Acids Res* **44**, 5732-5742 (2016).

30. M. Falkenberg *et al.*, Mitochondrial transcription factors B1 and B2 activate transcription of human mtDNA. *Nat Genet* **31**, 289-294 (2002).

31. T. J. Nicholls, M. Minczuk, In D-loop: 40 years of mitochondrial 7S DNA. *Exp Gerontol* **56**, 175-181 (2014).

32. H. Ma, P. H. O'Farrell, Selfish drive can trump function when animal mitochondrial genomes compete. *Nat Genet* **48**, 798-802 (2016).

33. H. Ma *et al.*, Metabolic rescue in pluripotent cells from patients with mtDNA disease. *Nature* **524**, 234-238 (2015).

34. L. A. Hyslop *et al.*, Towards clinical application of pronuclear transfer to prevent mitochondrial DNA disease. *Nature* **534**, 383-386 (2016).

35. M. Yamada *et al.*, Genetic Drift Can Compromise Mitochondrial Replacement by Nuclear Transfer in Human Oocytes. *Cell stem cell* **18**, 749-754 (2016).

36. E. Kang *et al.*, Mitochondrial replacement in human oocytes carrying pathogenic mitochondrial DNA mutations. *Nature* **540**, 270-275 (2016).

37. C. W. Birky, Relaxed and stringent genomes: why cytoplasmic genes don't obey Mendel's laws. *J Heredity* **85**, 355-365 (1994).

38. P. F. Chinnery, D. C. Samuels, Relaxed replication of mtDNA: a model with implications for the expression of disease. *Am J Hum Genet* **64**, 1158-1165 (1999).

39. P. F. Chinnery, D. C. Samuels, J. Elson, D. M. Turnbull, Accumulation of mitochondrial DNA mutations in ageing, cancer, and mitochondrial disease: is there a common mechanism? *Lancet* **360**, 1323-1325 (2002).

40. V. I. Floros *et al.*, Segregation of mitochondrial DNA heteroplasmy through a developmental genetic bottleneck in human embryos. *Nat Cell Biol* **20**, 144-151 (2018).

41. M. Ginsburg, M. H. L. Snow, A. McLaren, Primordial germ cells in the mouse embryo during gastrulation. *Development* **110**, 521-528 (1990).

42. W. J. Kent *et al.*, The human genome browser at UCSC. *Genome Research* **12**, 996-1006 (2002).

43. M. D. Hendy, M. D. Woodhams, A. Dodd, Modelling mitochondrial site polymorphisms to infer the number of segregating units and mutation rate. *Biol Lett* **5**, 397-400 (2009).

44. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993 (2011).

45. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

46. A. Tan, G. R. Abecasis, H. M. Kang, Unified representation of genetic variants. *Bioinformatics* **31**, 2202-2204 (2015).

47. W. McLaren *et al.*, The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).

48. G. Jun *et al.*, Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* **91**, 839-848 (2012).

49. R. M. Andrews *et al.*, Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* **23**, 147 (1999).

50. C. Genomes Project *et al.*, A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

51. S. Purcell *et al.*, PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007).

52. A. Manichaikul *et al.*, Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873 (2010).

53. X. Zheng *et al.*, A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326-3328 (2012).

54. M. P. Conomos, A. P. Reiner, B. S. Weir, T. A. Thornton, Model-free Estimation of Recent Genetic Relatedness. *Am J Hum Genet* **98**, 127-148 (2016).
55. M. P. T. Conomos, T.; Gogarten, S. M., GENESIS: GENetic EStimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness. *R package version 2.6.1.* (2017).
56. J. Staples, D. A. Nickerson, J. E. Below, Utilizing graph theory to select the largest set of unrelated individuals for genetic analysis. *Genet Epidemiol* **37**, 136-141 (2013).
57. H. Weissensteiner *et al.*, HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res* **44**, W58-63 (2016).
58. L. B. Alexandrov *et al.*, Signatures of mutational processes in human cancer. *Nature* **500**, 415-421 (2013).
59. M. Krzywinski *et al.*, Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645 (2009).

Acknowledgments: We gratefully acknowledge the patients, families, and health care professionals involved in the NIHR BioResource – Rare Diseases and the 100,000 Genomes projects. We are grateful to N. S. Jones for his critical comments on an early draft of the manuscript. **Funding:** This study makes use of data generated by the NIHR BioResource and the Genomics England Rare Diseases pilot projects. Genotype and phenotype data of both projects are part of the 100,000 Genomes Project. The main source of funding for the BioResource and Genomics England is provided by the National Institute for Health Research of England (NIHR, http://www.nihr.ac.uk). This work was also made possible by funding from the UK Medical Research Council (MRC) to create the UK Clinical Genomics Datacentre. PFC is a Wellcome Trust Principal Research Fellow (101876/Z/13/Z & 212219/Z/18/Z), and a NIHR Senior Investigator, who receives support from the Medical Research Council Mitochondrial Biology Unit (MC_UP_1501/2), the Evelyn Trust, and the NIHR Biomedical Research Centre based at Cambridge University Hospitals NHS Foundation Trust and the University of Cambridge. WHO is a NIHR Senior Investigator and his laboratory receives support from the British Heart Foundation, Bristol-Myers Squibb, European Commission, MRC, NHS Blood and Transplant, Rosetrees Trust, and the NIHR Biomedical Research Centre based at Cambridge University Hospitals NHS Foundation Trust and the University of Cambridge. MC is an NIHR Senior Investigator and is funded by the NIHR Biomedical Research Centre at St Bartholomew's Hospital. JeT, JoT, SP and AOMW are funded by the NIHR Biomedical Research Centre, Oxford. This work was supported in part by Wellcome Trust grant 090532/Z/09/Z. RH is funded by Wellcome Trust grants 201064/Z/16/Z, 109915/Z/15/Z, 203105/Z/16/Z, MRC UK grant MR/N025431/1, ERC grant 309548 and Newton Fund MR/N027302/1. JS is funded by MRC UK grant MR/M012212/1. AM, GA and AW are funded by the Moorfields Eye Charity. GA and

AW are funded by the RP Fighting Blindness. All Moorfields Eye Hospital and Institute of Ophthamology authors are funded by the UCL Institute of Ophthalmology and Moorfields NIHR Biomedical Resource Centre. The Bristol NIHR Biomedical Research Centre provided infrastructure for BioResource activities in Bristol. Additional NIHR Biomedical Research Centres that contributed include Imperial College Healthcare NHS Trust BRC, Guy's and St Thomas' NHS Foundation Trust and King's College London BRC. The authors listed also represent NephroS, the UK study of Nephrotic Syndrome. AL is a British Heart Foundation Senior Basic Science Research Fellow - FS/13/48/30453. DLB, ACT, NVZ and MIM are members of the DOLORisk consortium funded by the European Commission Horizon 2020 (ID633491). ACT is a member of the International Diabetic Neuropathy Consortium, the Novo Nordisk Foundation (Ref. NNF14SA0006). DLB is a Wellcome clinical scientist (202747/Z/16/Z). ARW is supported by the NIHR-BRC of UCL Institute of Ophthalmology and Moorfields Eye Hospital. IR and EL are supported by the NIHR Translational Research Collaboration- Rare Diseases. HJB works for the Netherlands CardioVascular Research Initiative (CVON). TKB is sponsored by the NHSBT and British Society of Haematology. KGCS holds a Wellcome Investigator Award, MRC Programme Grant (number MR/L019027/1). MIM is a Wellcome Senior Investigator. Support from the Wellcome grant numbers 090532, 0938381. PHD receives funding from ICP Support. HSM receives support from BHF Programme Grant no. RG/16/4/32218. NC is partially funded by Imperial College NIHR BRC. MRW holds a NIHR award to the NIHR Imperial Clinical Research Facility at Imperial College Healthcare NHS Trust. PYWM is supported by grants from MRC UK (G1002570), Fight for Sight (1570/1571), Fight for Sight (24TP171), NIHR (IS-BRC-1215-20002). RH is a Wellcome Trust Investigator (109915/Z/15/Z), who receives support from the Wellcome Centre for

Mitochondrial Research (203105/Z/16/Z), Medical Research Council (UK) (MR/N025431/1), the European Research Council (309548), the Wellcome Trust Pathfinder Scheme (201064/Z/16/Z), the Newton Fund (UK/Turkey, MR/N027302/1) and the European Union H2020 – Research and Innovation Actions (SC1-PM-03-2017, Solve-RD). KF and CVG were supported by the Research Council of the University of Leuven (BOF KU Leuven, Belgium; OT/14/098). JSW is funded by Welcome Trust [107469/Z/15/Z]; (ii) National Institute for Health Research (NIHR) Cardiovascular Biomedical Research Unit at Royal Brompton & Harefield NHS Foundation Trust and Imperial College London. GA is funded by NIHR-Biomedical Research Centre at Moorfields Eye Hospital and UCL Institute of Ophthalmology, Fight for Sight (UK) Early Career Investigator Award, Moorfields Eye Hospital Special Trustees, Moorfields Eye Charity, Foundation Fighting Blindness (USA) and Retinitis Pigmentosa Fighting Blindness. MCS holds a MRC Clinical Research Training Fellowship, grant ref MR/R002363/1. MAKu holds a NIHR Research Professorship NIHR-RP-2016-07-019 and Wellcome Intermediate Fellowship 098524/Z/12/A. JWhi is a recipient of a Cancer Research UK Cambridge Cancer Centre Clinical Research Training Fellowship. AJM has received funding from a Medical Research Council Senior Clinical Fellowship (MR/L006340/1). DPG is funded by the MRC, Kidney Research UK and St Peters Trust for Kidney, Bladder and Prostate Research. SAJ is funded by Kids Kidney Research. CL received funding from a MRC Clinical Research Training Fellowship (MR/J011711/1). KD is a HSST trainee supported by Health Education England. CHad was funded through a PhD Fellowship by the NIHR Translational Research Collaboration Rare Diseases. MJD receives funding from Wellcome Trust (WT098519MA). KJM is supported by the Northern Counties Kidney Research Fund. ELM a proportion of my work was undertaken at University College London Hospitals/University

College London, which received a proportion of funding from the Department of Health's National Institute for Health Research Biomedical Research Centres funding scheme. KCG is a holder of NIHR – BRC funding. PLB is a NIHR Senior Investigator. This research was partly funded by the NIHR Great Ormond Street Hospital Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. **Author contributions:** Study design: PFC, FLR, MC, WHO, E., WW. Data analysis: ET, WW, ST, MJK. Writing: PFC, ET, WW. Experimental and analytical supervision: PFC and ET, Project Supervision: PFC, and ET. The remaining authors contributed to the recruitment of participants, sample logistics and initial data preparation. **Competing interests:** MIM: serves on advisory panels for Pfizer, NovoNordisk, Zoe Global; has received honoraria from Pfizer, NovoNordisk and Eli Lilly; has stock options in Zoe Global; has received research funding from Abbvie, Astra Zeneca, Boehringer Ingelheim, Eli Lilly, Janssen, Merck, NovoNordisk, Pfizer, Roche, Sanofi Aventis, Servier, Takeda. TJA has received consultancy payments from AstraZeneca within the last 5 years and has received speaker honoraria from Illumina. KJM previously received funding for research and currently on the scientific advisory board of Gemini Therapeutics, Boston, USA. MCS received travel and accommodation fees from NovoNordisk. DML serves on advisory boards for Agios, Novartis and Cerus. AMK had no competing interests at the time of the study, since the study has received an educational grant from CSL Behring to attend the ISTH meeting in Berlin in 2017. CVG is holder of the Bayer and Norbert Heimburger (CSL Behring) Chairs. **Data and materials availability:** Heteroplasmy data for the mother-child pairs is provided in the external database S1 (Data S1). Whole genome sequence data from the NIHR BioResource – Rare Diseases project can be found in the European Genome-phenome Archive (EGA) at the EMBL European Bioinformatics Institute

(BPD: EGAD00001004519, CSVD: EGAD00001004513, HCM: EGAD00001004514, ICP: EGAD00001004515, IRD: EGAD00001004520, MPMT: EGAD00001004521, NDD: EGAD00001004522, NPD: EGAD00001004516, PAH: EGAD00001004525, PID: EGAD00001004523, PMG: EGAD00001004517, SMD: EGAD00001004524, SRNS: EGAD00001004518, See **Table S5** for the disease abbreviations). Whole genome sequence data from the UK Biobank samples are available through a data release process overseen by UK Biobank (https://www.ukbiobank.ac.uk/). Whole genome sequence data from the participants enrolled in 100,000 Genomes Project can be accessed via Genomics England Limited following the procedure outlined at: https://www.genomicsengland.co.uk/about-gecip/joining-research-community/

Supplementary Materials:

Materials and Methods

Extended authors and all author addresses and affiliations

Figures S1-S15

Tables S1-S5

External Database S1

References (*44-59*)

Figures

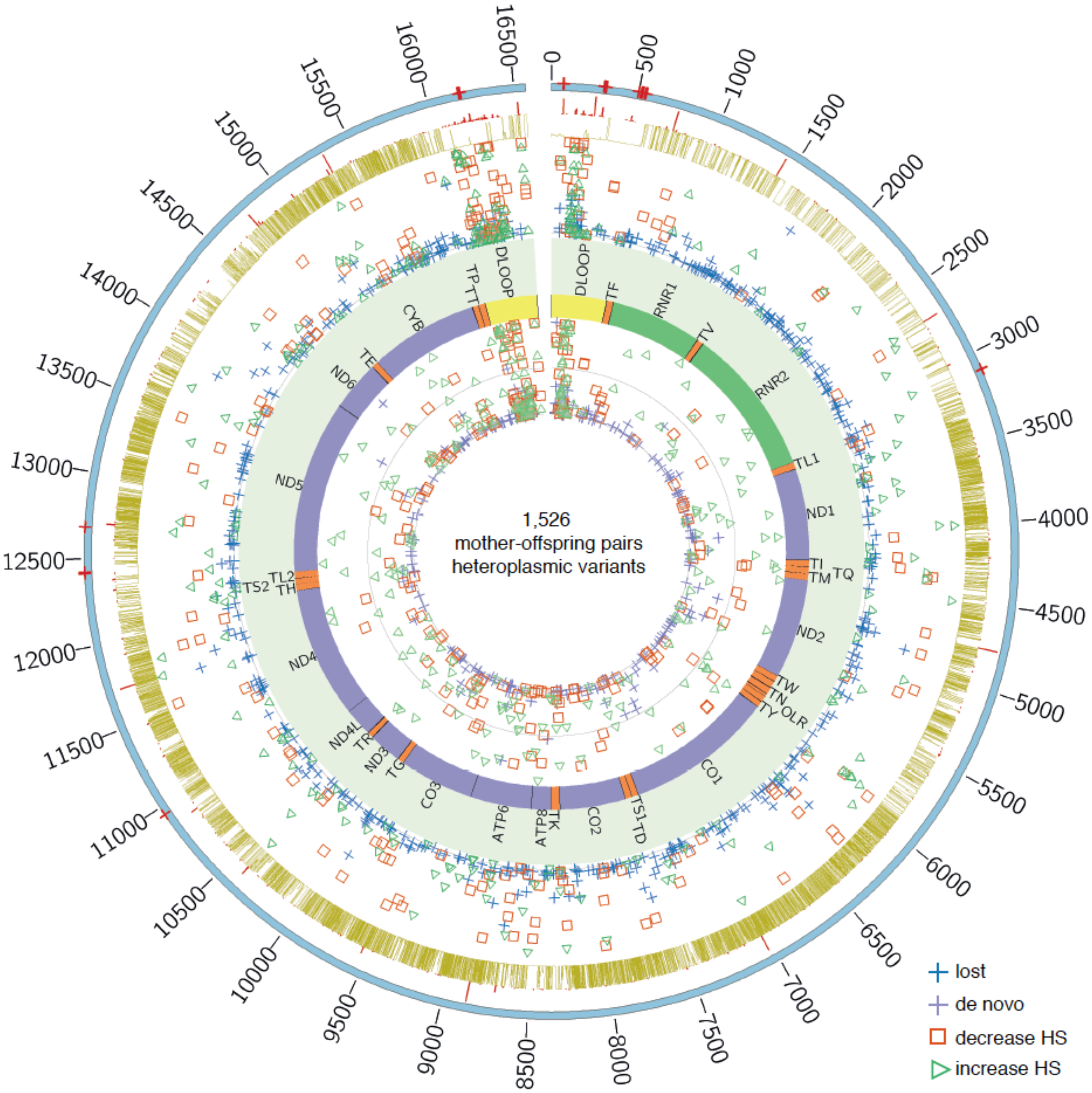

0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
7000
7500
8000
8500
9000
9500
10000
10500
11000
11500
12000
12500
13000
13500
14000
14500
15000
15500
16000
16500
DLOOP
TF
RNR1
TV
RNR2
TL1
ND1
TI
TM
TQ
ND2
TW
TN
TY
OLR
CO1
TS1-TD
CO2
TK
ATP8
ATP6
CO3
TG
ND3
TR
ND4L
ND4
TL2
TH
TS2
ND5
ND6
TE
CYB
TP-TT
DLOOP
1,526
mother-offspring pairs
heteroplasmic variants
+ lost
+ de novo
□ decrease HS
▷ increase HS

Fig 1. Circos plot of mitochondrial heteroplasmic variants identified in 1,526 mother-offspring pairs.

Circles from the outside to the inside indicate the following: (1) position of a variant on the mtDNA, the removed regions are shown in red crosses; (2) Minor allele frequency for common variants (MAF>1%) derived from 30,506 NCBI mtDNA sequences (*5*), where the radial axis corresponds to the MAF; (3) phastCons100way scores from UCSC (*42*) where the radial axis corresponds to the degree of conservation ; (4) heteroplasmic variants identified in the mothers where the radial axis corresponds to the heteroplasmy fraction, HF; (5) regions corresponding to the different mtDNA genes (yellow - D-loop, purple – coding region, green – rRNAs and orange - tRNAs); (6) heteroplasmic variants identified in the offspring where the radial axis corresponds to the heteroplasmy fraction, HF.

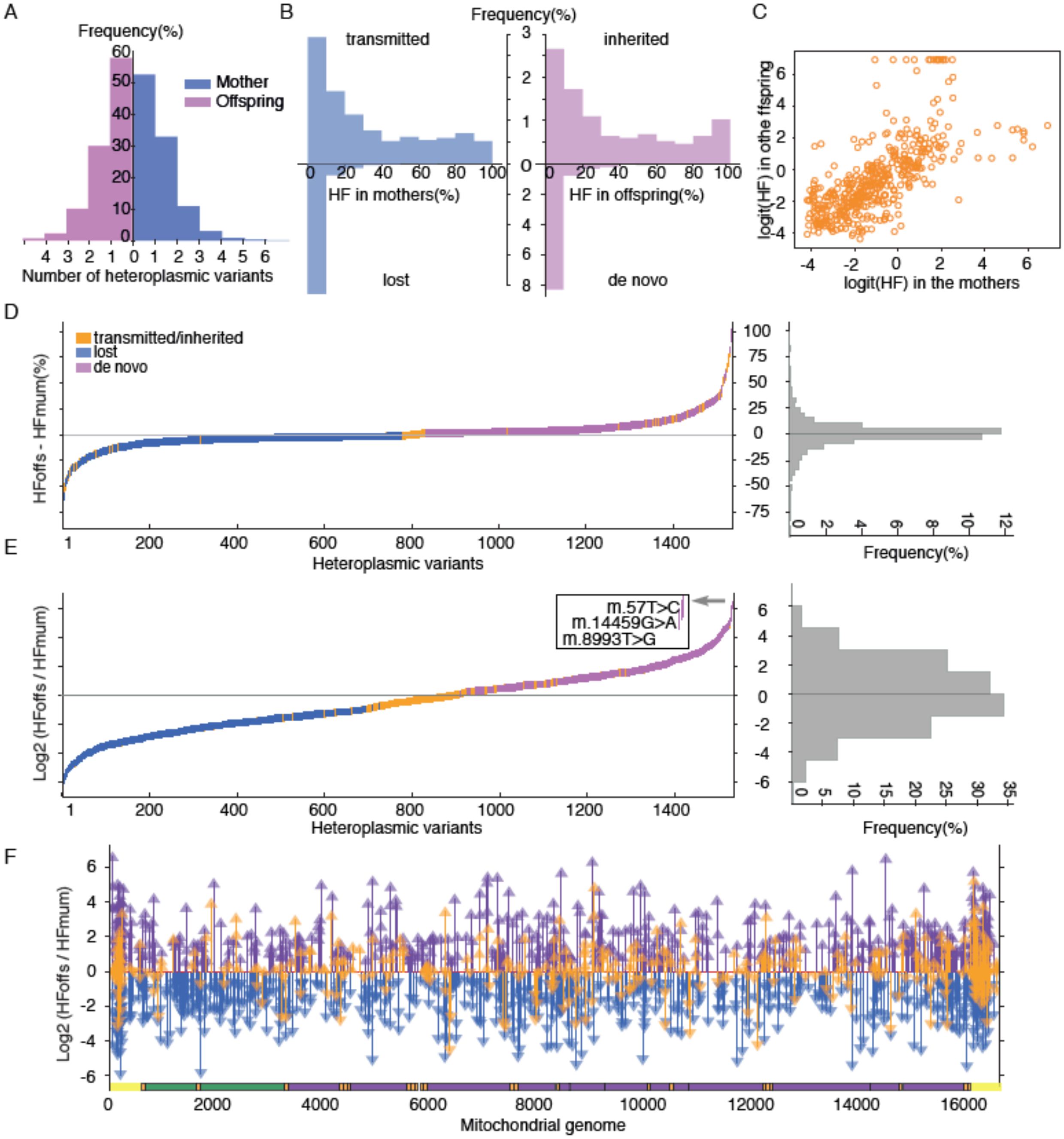

A
Frequency(%)
60
50
40
30
20
10
0
Mother
Offspring
4 3 2 1 0 0 1 2 3 4 5 6
Number of heteroplasmic variants
B
Frequency(%)
transmitted
inherited
3
2
1
0
0 20 40 60 80 100
HF in mothers(%)
0 20 40 60 80 100
HF in offspring(%)
2
4
6
8
lost
de novo
C
logit(HF) in othe ffspring
6
4
2
0
-2
-4
-4 -2 0 2 4 6
logit(HF) in the mothers
D
transmitted/inherited
lost
de novo
HFoffs - HFmum(%)
1 200 400 600 800 1000 1200 1400
Heteroplasmic variants
100
75
50
25
0
-25
-50
-75
0 2 4 6 8 10 12
Frequency(%)
E
Log2 (HFoffs / HFmum)
m.57T>C
m.14459G>A
m.8993T>G
1 200 400 600 800 1000 1200 1400
Heteroplasmic variants
6
4
2
0
-2
-4
-6
0 5 10 15 20 25 30 35
Frequency(%)
F
Log2 (HFoffs / HFmum)
6
4
2
0
-2
-4
-6
0 2000 4000 6000 8000 10000 12000 14000 16000
Mitochondrial genome

Fig 2. Transmission of heteroplasmic mtDNA variants in 1,562 mother-offspring pairs.

(**A**) Frequency distribution of number of heteroplasmic variants in the mothers and offspring. (**B**) Distribution of HF in the mothers and offspring, transmitted, inherited, lost and *de novo* are shown separately. (**C**) Scatter plot of logit(HF) in transmitted heteroplasmic variants between the mothers and offspring (R^2=0.79, P=1.52 x 10^{-93}, Pearson's correlation). (**D**) Left, difference in the percentage shift of HF between the offspring and the corresponding mothers (HFoffspring - HFmother) ordered by the degree of shift. Right, distribution of the difference of the percentage shift of HF between offspring and the corresponding mothers (HFoffspring - HFmother). (**E**) Left, log2 ratio of HF difference between offspring and the corresponding mothers ordered by the degree of log2 ratio. Three increase HSs with values above 6 shown in the box. Right, distribution of log2 ratio of HF difference between offspring and the corresponding mothers. (**F**) log2 ratio of HF difference between offspring and the corresponding mothers aligned to the whole mitochondrial DNA sequence; the mtDNA regions are shown at the bottom bar in different colors (yellow - D-loop, purple – coding region, green – rRNAs and orange - tRNAs).

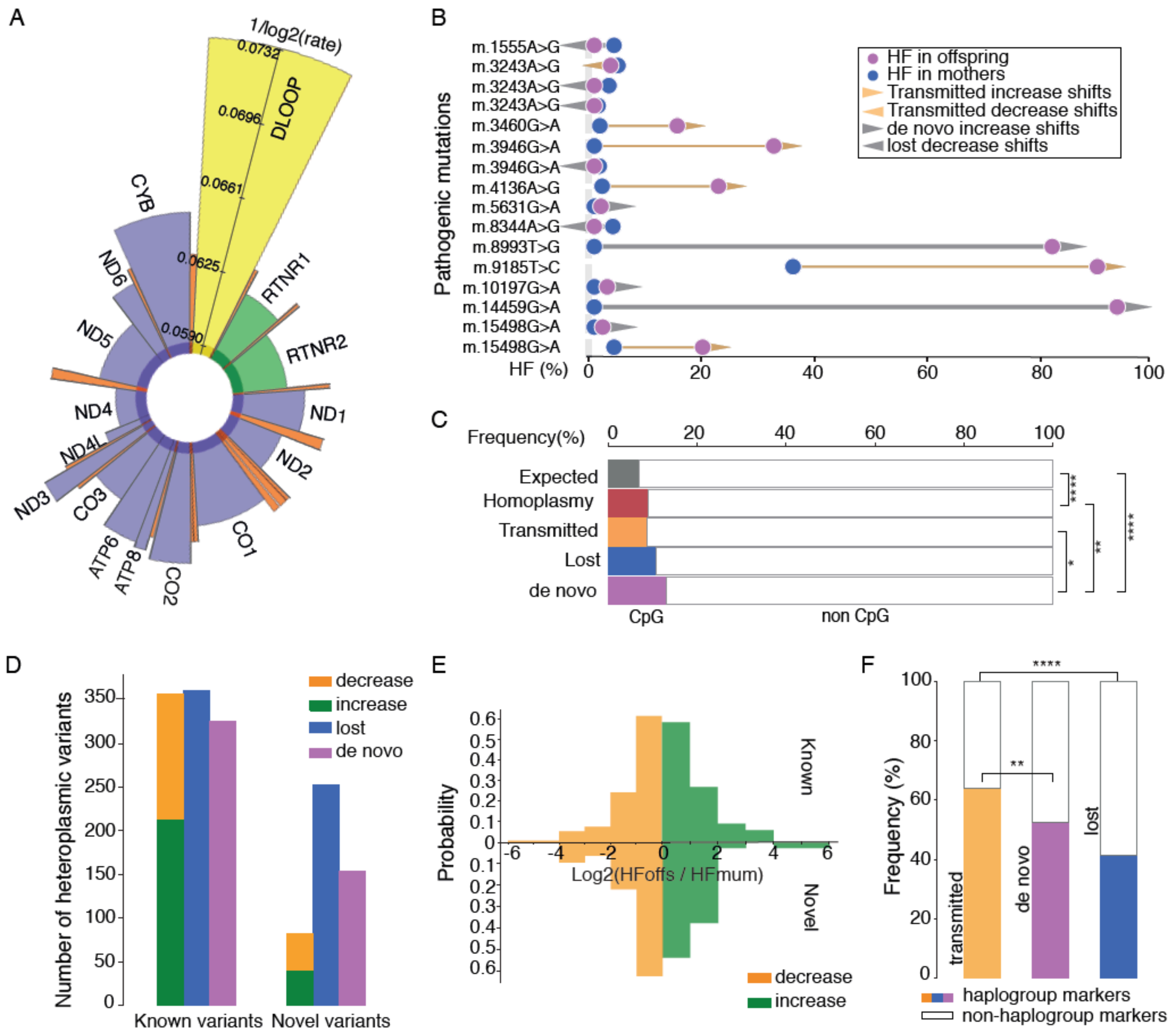

A
1/log2(rate)
0.0732
0.0696
0.0661
0.0625
0.0590
DLOOP
CYB
ND6
ND5
ND4
ND4L
ND3
CO3
ATP6
ATP8
CO2
CO1
ND2
ND1
RTNR2
RTNR1
B
Pathogenic mutations
m.1555A>G
m.3243A>G
m.3243A>G
m.3243A>G
m.3460G>A
m.3946G>A
m.3946G>A
m.4136A>G
m.5631G>A
m.8344A>G
m.8993T>G
m.9185T>C
m.10197G>A
m.14459G>A
m.15498G>A
m.15498G>A
HF (%)
0
20
40
60
80
100
HF in offspring
HF in mothers
Transmitted increase shifts
Transmitted decrease shifts
de novo increase shifts
lost decrease shifts
C
Frequency(%)
0
20
40
60
80
100
Expected
Homoplasmy
Transmitted
Lost
de novo
CpG
non CpG
****
**
*
****
D
Number of heteroplasmic variants
0
50
100
150
200
250
300
350
decrease
increase
lost
de novo
Known variants
Novel variants
E
Probability
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
-6
-4
-2
0
2
4
6
Log2(HFoffs / HFmum)
Known
Novel
decrease
increase
F
Frequency (%)
0
20
40
60
80
100
****
**
transmitted
de novo
lost
haplogroup markers
non-haplogroup markers

Fig 3. Characteristics of the heteroplasmic mtDNA variants in 1,562 mother-offspring pairs.

(**A**) Mutation rate of mtDNA genomic regions was estimated using 477 *de novo* heteroplasmic variants from 1,526 mother-offspring pairs detected at HF>1%. Vertical axes represent 1/log2(mutation rate) per base per mother-child transmission. mtDNA genomic regions are labeled and shown in different colors (yellow - D-loop, purple – coding region, green – rRNAs and orange - tRNAs). All tRNAs were combined to estimate the tRNA mutation rate. Note that this is the raw number of new mutations/bp/transmission detected at HF>1% in the offspring, and does not factor in the detection threshold nor segregation because current models assume neutrality(*13, 43*), which we later show is not the case. (**B**) Pathogenic mutations were observed in 1,526 mother-offspring pairs. Each dot represents the HF in the mothers (blue) and the corresponding offspring (pink); the directions of arrow show increase (->) or decrease (<-) HS; the length of the arrow between each pair of points represents the change in HF (orange - transmitted heteroplasmic variants, grey - de novo / lost heteroplasmic variants). (**C**) Frequency of heteroplasmic variants at CpG and non-CpG islands. Expected, homplasmic variants, transmitted, lost and *de novo* heteroplasmic variants are shown separately. (**D**) Number of novel versus known heteroplasmic variants in transmitted, lost and *de novo* heteroplasmic variants, increase and decreasing HS in transmitted heteroplasmic variants are shown in different colors. (**E**) Distribution of HS between the offspring and the corresponding mothers in transmitted known and novel heteroplasmic variants, increase and decreasing HS are shown in different colors. (**F**) Frequency of haplogroup defining variants in transmitted, lost and de novo heteroplasmic variants. The transmitted heteroplasmic variants were more likely to affect known haplogroup specific sites on the world mtDNA phylogeny than the lost and *de novo* heteroplasmic variants (*P*=7.86×10^{-11} and *P*=0.0016 respectively, Fisher's exact test). P value < 0.05*, < 0.01**, < 0.001*** and <0.0001****.

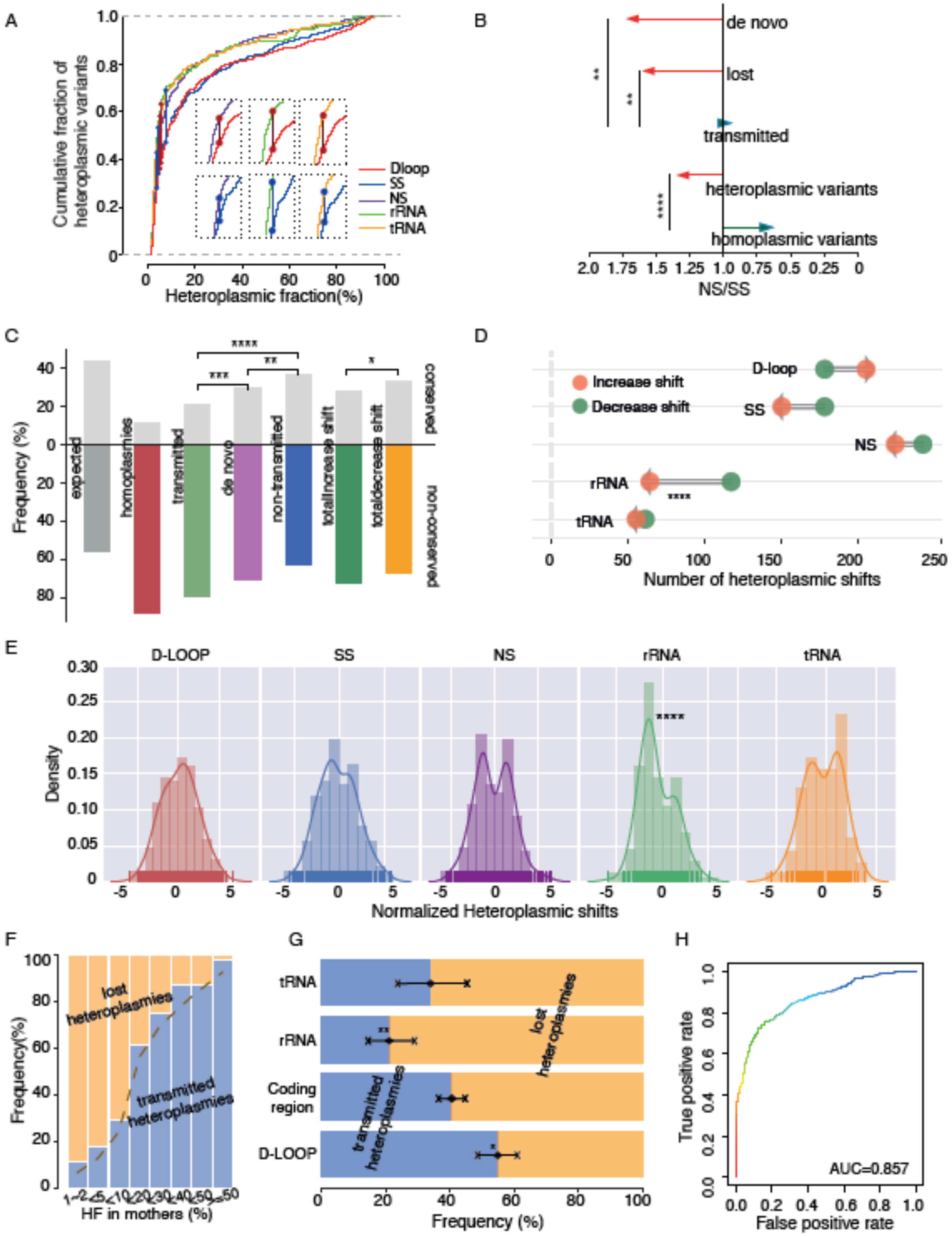
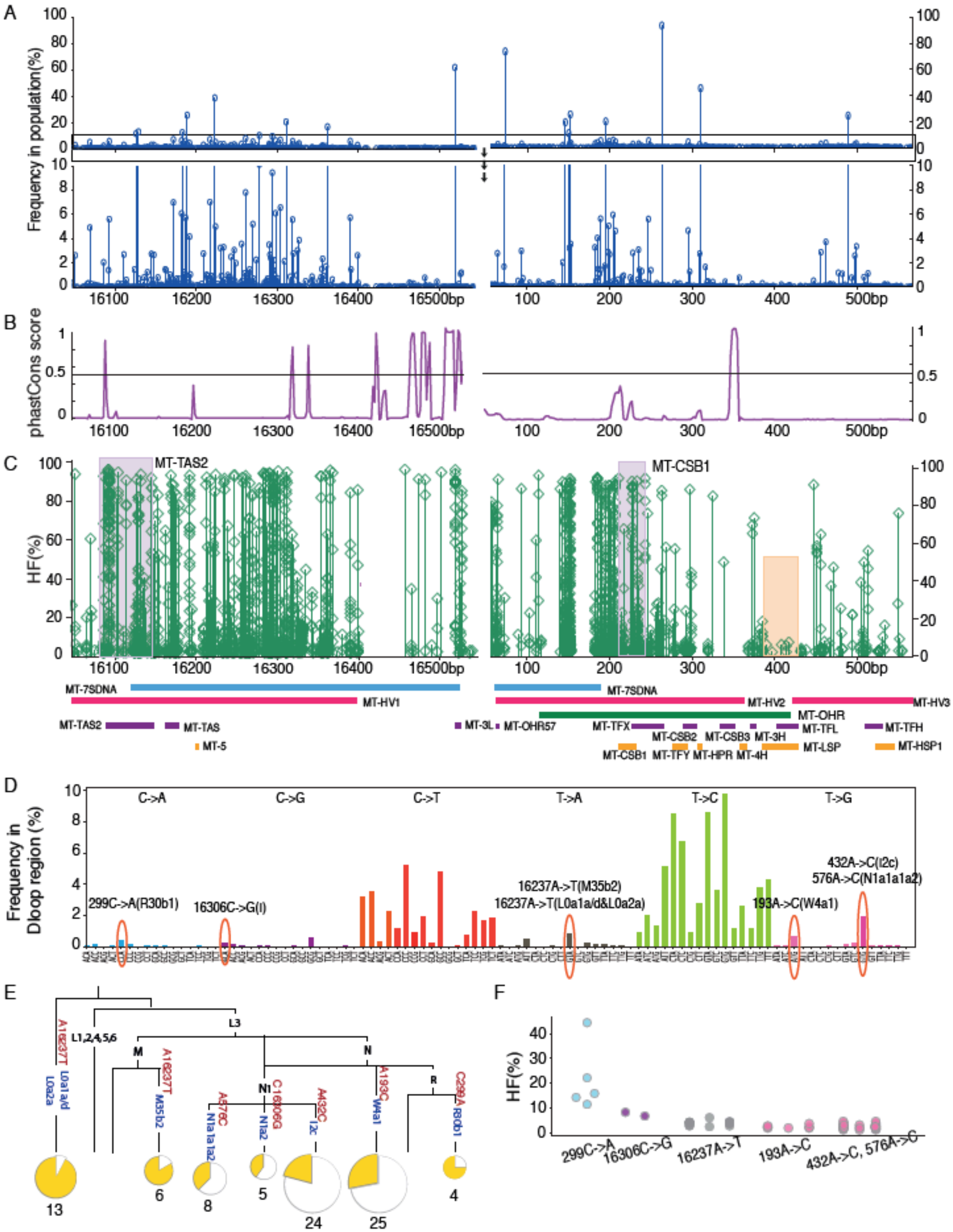

A
1.0
0.8
0.6
0.4
0.2
0
Cumulative fraction of heteroplasmic variants
0
20
40
60
80
100
Heteroplasmic fraction(%)
Dloop
SS
NS
rRNA
tRNA
B
de novo
lost
transmitted
heteroplasmic variants
homoplasmic variants
**
**
****
2.0
1.75
1.5
1.25
1.0
0.75
0.5
0.25
0
NS/SS
C
Frequency (%)
40
20
0
20
40
60
80
expected
homoplasmies
transmitted
de novo
non-transmitted
totalincrease shift
totaldecrease shift
****
***
**
*
conserved
non-conserved
D
Increase shift
Decrease shift
D-loop
SS
NS
rRNA
tRNA
****
0
50
100
150
200
250
Number of heteroplasmic shifts
E
D-LOOP
SS
NS
rRNA
tRNA
****
0.30
0.25
0.20
0.15
0.10
0.05
0
Density
-5
0
5
Normalized Heteroplasmic shifts
F
100
80
60
40
20
0
Frequency(%)
lost heteroplasmies
transmitted heteroplasmies
1~2
<5
<10
<20
<30
<40
<50
>=50
HF in mothers (%)
G
tRNA
rRNA
Coding region
D-LOOP
**
*
lost heteroplasmies
transmitted heteroplasmies
0
20
40
60
80
100
Frequency (%)
H
1.0
0.8
0.6
0.4
0.2
0.0
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
AUC=0.857

Fig 4. Evidence of selection during the transmission of mtDNA heteroplasmy in 1,526 mother-offspring pairs.

(**A**) Cumulative distributions of HF in the mothers and offspring within each mtDNA region. Vertical lines between two curves show the greatest distance between Dloop / SS and NS / rRNA / tRNA regions (*P*-values in **table S3**). (**B**) NS/SS ratio of NS and SS variants for observed homoplasmic polymorphisms, total heteroplasmic variants, transmitted, lost and *de novo* heteroplasmic variants. (**C**) Frequency of heteroplasmic variants affecting conserved and non-conserved sites. Expected, homoplasmic variants, transmitted, lost, *de novo* heteroplasmic variants, increase and decrease HSs are shown separately. (**D**) Number of heteroplasmies showing an increase or decrease HF in each mtDNA region. Left-facing arrows indicate that the number increasing was less than the number decreasing. Right-facing arrows indicate that the number increasing was greater than the number decreasing. (**E**) Histograms of HS in each mtDNA region with fitted kernel density curves. (**F**) Bar plot of the frequency of transmitted heteroplasmic variants by bins of HF in the mothers. (**G**) Frequency of transmitted heteroplasmic variants in each mtDNA region, along with 95% confidence intervals. (**H**) Receiver operating characteristic (ROC) curve for the logistic regression model of transmission (Area under the curve, AUC=0.857). P value < 0.05*, < 0.01**, < 0.001*** and <0.0001****.



A
Frequency in population(%)
100
80
60
40
20
0
10
8
6
4
2
0
16100
16200
16300
16400
16500bp
100
200
300
400
500bp
B
phastCons score
1
0.5
0
C
HF(%)
MT-TAS2
MT-CSB1
MT-7SDNA
MT-HV1
MT-HV2
MT-HV3
MT-OHR
MT-TAS2
MT-TAS
MT-5
MT-3L
MT-OHR57
MT-TFX
MT-CSB2
MT-CSB3
MT-3H
MT-TFL
MT-TFH
MT-CSB1
MT-TFY
MT-HPR
MT-4H
MT-LSP
MT-HSP1
D
Frequency in Dloop region (%)
C->A
C->G
C->T
T->A
T->C
T->G
299C->A(R30b1)
16306C->G(I)
16237A->T(M35b2)
16237A->T(L0a1a/d&L0a2a)
193A->C(W4a1)
432A->C(I2c)
576A->C(N1a1a1a2)
E
L3
L1,2,4,5,6
M
N
N1
R
A16237T
L0a1a/d
L0a2a
A16237T
M35b2
A576C
N1a1a1a2
C16306G
N1a2
A432C
I2c
A193C
W4a1
C299A
R30b1
13
6
8
5
24
25
4
F
HF(%)
40
30
20
10
0
299C->A
16306C->G
16237A->T
193A->C
432A->C, 576A->C

Fig 5. The distribution of heteroplasmic variants in mtDNA Dloop region.

(**A**) MAF of homoplasmic single nucleotide polymorphisms observed in 30,506 NCBI mtDNA sequences, with an expanded axis to show MAF<10% at the bottom. (**B**) Trend of PhastCons scores is shown across the mtDNA D-loop region. (**C**) HFs observed in 12,975 mtDNA sequences in the D-loop region. MT-TAS2 and MT-CSB1 are shadowed in light purple. MT-LSP is shadowed in light orange. Corresponding known sub-regions of the mtDNA D-loop are shown at the bottom. Key - MT-3H: mt3 H-strand control element, MT-3L: L-strand control element, MT-4H: mt4 H-strand control element, MT-7SDNA: 7S DNA, MT-CSB1: Conserved sequence block 1, MT-CSB2: Conserved sequence block 2, MT-CSB3: Conserved sequence block 3, MT-HPR: replication primer, MT-HSP1: Major H-strand promoter, MT-HV1: Hypervariable segment 1, MT-HV2: Hypervariable segment 2, MT-HV3: Hypervariable segment 3, MT-LSP: L-strand promoter, MT-OHR: H-strand origins, MT-OHR57: H-strand origin, MT-TAS: termination-associated sequence, MT-TAS2: extended termination-associated sequence, MT-TFH/MT-TFL/ MT-TFX/ MT-TFY/: mtTF1 binding site. (**D**) Trinucleotide mutational signature of heteroplasmic variants in the D-loop region in 12,975 mtDNA sequences. The bars representing the frequency for the six types of substitution are displayed in different colours. Labeled heteroplasmic variants are included in the bars with the red circles. (**E**) Simplified mtDNA phylogeny tree showing 6 heteroplasmic variants (refer to main text). Variants are shown in red and haplogroups are shown in blue. The pie chart sizes are proportional to the number of samples (shown at the bottom) belonging to the corresponding haplogroup in 10,210 unrelated mtDNA sequences. The proportion of samples carrying each heteroplasmic variant within the same haplogroup is shown in yellow. (**F**) HF of six heteroplasmic variants shared by more than one individual who belong to the same haplogroup.

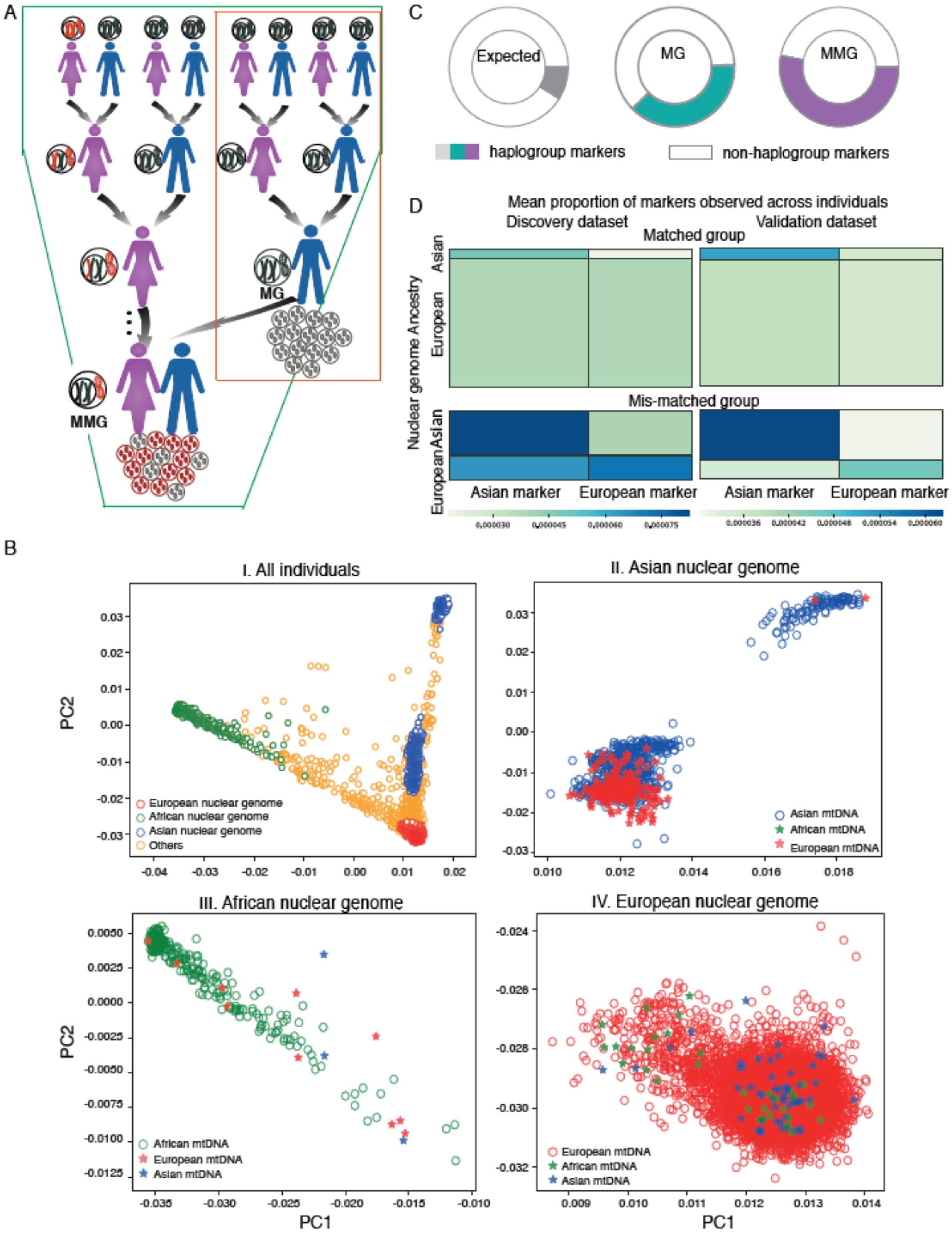

A
MG
MMG
C
Expected
MG
MMG
haplogroup markers
non-haplogroup markers
D
Mean proportion of markers observed across individuals
Discovery dataset
Validation dataset
Matched group
Mis-matched group
Nuclear genome Ancestry
Asian
European
Asian
European
Asian marker
European marker
Asian marker
European marker
0.000030
0.000045
0.000060
0.000075
0.000036
0.000042
0.000048
0.000054
0.000060
B
I. All individuals
PC2
0.03
0.02
0.01
0.00
-0.01
-0.02
-0.03
-0.04
-0.03
-0.02
-0.01
0.00
0.01
0.02
European nuclear genome
African nuclear genome
Asian nuclear genome
Others
II. Asian nuclear genome
0.03
0.02
0.01
0.00
-0.01
-0.02
-0.03
0.010
0.012
0.014
0.016
0.018
Asian mtDNA
African mtDNA
European mtDNA
III. African nuclear genome
PC2
0.0050
0.0025
0.0000
-0.0025
-0.0050
-0.0075
-0.0100
-0.0125
-0.035
-0.030
-0.025
-0.020
-0.015
-0.010
PC1
African mtDNA
European mtDNA
Asian mtDNA
IV. European nuclear genome
-0.024
-0.026
-0.028
-0.030
-0.032
0.009
0.010
0.011
0.012
0.013
0.014
PC1
European mtDNA
African mtDNA
Asian mtDNA

Fig 6. The characteristics of heteroplasmic variants in the nuclear ancestry and mtDNA ancestry matched and mismatched groups.

(**A**) Schematic showing how individuals with matched (MG, red border) and mismatched (MMG, green border) nuclear and mtDNA genomes arise over generations. Red and grey colored mtDNAs represent two different hypothetical populations. (**B**) I. Projection of the nuclear genotypes at common SNPs onto the two leading principal components computed with the 1000 Genomes dataset, with individuals colored by their assigned nuclear ancestry: Asian (blue), African (green), European (red) and Other (orange). The individuals colored in blue, green and red in the boxes labelled II, III and IV are shown in panels II, III and IV, respectively, where they are colored by their mitochondrial ancestries. Stars indicate that the mitochondrial ancestry does not match the nuclear ancestry. (**C**) Proportion of haplogroup defining variants in the matched group (MG) and mismatched group (MMG) in 9,631 mtDNA sequences from unrelated individuals, along with the expected proportion shown at the left side. Distinct heteroplasmic sites were more likely to affect known haplogroup specific sites (*26*) than the rest of the mitochondrial genome compared to that expected by chance ($P<2.2 \times 10^{-16}$, Fisher's exact test). This bias was stronger in the mismatched group than the matched group ($P=0.001$, Fisher's exact test). (**D**) Heatmaps showing the density of observed heteroplasmic mtDNA haplogroup-defining variants in the observation dataset (left) and validation dataset (right). The matched (top) and mismatched (bottom) groups are shown separately, broken down by the nuclear ancestry of the carrier and the major haplogroup of the variants. The width of each column is proportional to the number of variants defining each of the two major haplogroups (Asian and European). Within each heatmap, the height of each row is proportional to the number of individuals having each nuclear ancestry. The density of heteroplasmic variants in each cell determines its color.

Supplementary Materials for

Germline selection shapes human mitochondrial DNA diversity*

Wei Wei, Salih Tuna, Michael J Keogh, Katherine R Smith[†], Timothy J Aitman, Phil L Beales, David L Bennett, Daniel P Gale, Maria A K Bitner-Glindzicz[#], Graeme C Black, Paul Brennan, Perry Elliott, Frances A Flinter, R Andres Floto, Henry Houlden, Melita Irving, Ania Koziell, Eamonn R Maher, Hugh S Markus, Nicholas W Morrell, William G Newman, Irene Roberts, John A Sayer, Kenneth G C Smith, Jenny C Taylor, Hugh Watkins, Andrew R Webster, Andrew O M Wilkie, Catherine Williamson, on behalf of the NIHR BioResource - Rare Diseases[+] and the 100,000 Genomes Project - Rare Diseases Pilot[+], Sofie Ashford, Christopher J Penkett, Kathleen E Stirrups, Augusto Rendon[†], Willem H Ouwehand[¶], John R Bradley[¶], F Lucy Raymond[¶], Mark Caulfield[†], Ernest Turro[§], Patrick F Chinnery[§]

[§]Correspondence to: pfc25@cam.ac.uk & et341@cam.ac.uk

[+] a full list of the authors is included in the Supplementary Material under ‘Extended Authors’
[¶] = equal contribution
[†] = equal contribution

# = deceased
§ = corresponding authors

* This manuscript has been accepted for publication in Science. This version has not undergone final editing. Please refer to the complete version of record at http://www.sciencemag.org/. The manuscript may not be reproduced or used in any manner that does not fall within the fair use provisions of the Copyright Act without the prior, written permission of AAAS.

This PDF file includes:

Materials and Methods
Extended authors and all author addresses and affiliations

Figs. S1-S15
Tables S1-S5
References (*44-59*)

Other Supplementary Materials for this manuscript include the following:

Data S1. List of heteroplasmic variants detected in 1,526 mother-offspring pairs

Materials and Methods

Study samples

Our primary data source was whole genome sequencing (WGS) data from whole-blood DNA from 13,037 individuals in the NIHR BioResource - Rare Diseases and 100,000 Genomes Project Pilot studies, including the patients affected by rare diseases and their relatives across several different disease domains (All the individuals are also included (**table S5**) (*22*). After quality control (QC) (details below), 12,975 samples including 1,526 mother-offspring pairs were included in this study. The ages of the mothers at sample collection ranged from 22 to 95 y (88 missing), the ages of the mothers at the time of giving birth ranged from 16 to 54 y (at least one person missing age in 89 pairs), and the ages of the offspring ranged from 0 to 48 y (88 missing).

Ethical approval

Ethical approval was provided by the East of England Cambridge South national research ethics committee (REC) under reference number: 13/EE/0325, with the majority of UK participants providing written informed consent for this approved study. Some UK participants were consented under an alternative REC-approved study obtained by the local investigators. Consent for non-UK samples was the responsibility of the respective principal investigators at the enrolling hospitals.

Extracting mitochondrial sequences and detecting variants

WGS was performed using the Illumina TruSeq DNA PCR-Free sample preparation kit (Illumina, Inc.) and an Illumina HiSeq 2500 sequencer, generating a mean depth of 45x (range from 34x to 72x) and greater than 15x for at least 95% of the reference human genome (**fig. S8A**).

The WGS reads were aligned to the Genome Reference Consortium human genome build 37 (GRCh37) using Isaac Genome Alignment Software (version 01.14; Illumina, Inc.) with options --cleanup-intermediary 1 --memory-limit 123 --base-quality-cutoff 15 --gap-scoring bwa --variable-read-length yes --ignore-missing-bcls 1 --ignore-missing-filters 1 --split-gap-length 10000 --per-tile-tls 1 --use-bases-mask Y150N1,Y150N1 --tiles s_4 --base-calls-format bcl-gz. The subset of sequencing reads which aligned to the mitochondrial genome were extracted from each sample's BAM file. We then ran MToolBox (v1.0) on the resulting smaller BAM files using the options mtdb_fasta=chrMT.fa, hg19_fasta=hg19RCRS.fa, mtdb=chrM, humandb=hg19RCRS, ref=RCRS, UseMarkDuplicates=true, UseIndelRealigner=true and MitoExtraction=true (*8*). MToolBox realigns the reads with GSNAP, which outputs multiple alignments per read up to a certain penalty score. It then uses mapExome.py to discard any read pairs for which either of the two reads in the pair mapped to multiple locations, including, for example, nuclear mitochondrial sequences in the nuclear genome and amplification artifacts. The VCFs were individually normalised and merged across all samples with bcftools (*44, 45*). The merged VCF was re-normalised with bcftools and vt (*46*) and duplicated variants were dropped with vt. The final VCF was annotated using the Variant Effect Predictor (VEP) (*47*).

Quality control of samples

Firstly, potential DNA cross-contamination was investigated using the nuclear genome. All samples passed contamination quality checks conducted by the sequencing provider Illumina, Inc. Additionally, we estimated the degree to which a DNA sample was contaminated by any other DNA sample using verifyBamID (*48*). Samples with an estimate of contamination (FREEMIX) exceeding 3% were excluded. The median FREEMIX value of the 13,037 samples included in the study was 0.042% and 99.84% of the samples had a FREEMIX value smaller than 1%.

Potential DNA cross-contamination was also checked using mtDNA variant calls. Samples carrying more than 20 heteroplasmic variants with HF >2% or carrying more than 5 heteroplasmies with similar HFs defining alternative haplogroups were excluded from this study. Additionally, we removed samples with the mean depth of the mtDNA genome < 200x. These QC checks removed 62 samples from the initial 13,037 mtDNA sequences. 12,975 individuals including 1,526 mother-offspring pairs were included to perform the analyses. We did not find an association between the logit (FREEMIX) parameter and presence of at least one mitochondrial heteroplasmy (*P*=0.89, logistic regression), further indicating that heteroplasmic calls were unlikely to be due to cross-sample contamination.

Quality control of variants

To define heteroplasmic variants, we selected variants as follows: 1) retain variants for which the confidence interval of VAF overlaps the range 1% to 98%; 2) remove all indels, as detecting heteroplasmic indels is unreliable using short-read sequencing data; 3) remove variants at sites with multiple alternate alleles, all of which have HFs < 5%; 4) remove variants falling within specific regions associated with misalignment errors related to homopolymeric tracts (np 66-71, 300-316, 513-525, 3106-3107, 12418-12425 and 16182-16194) (*49*); 5) remove heteroplasmic variants present in more than 2% of the individuals but not reported or vary rare in heteroplasmic variants in the literature (*14*) and 1000 Genomes individuals (np 499, 538, 545, 10953 and 12684); 6) remove variants less than 2 reads on each strand with the minor allele; 7) remove heteroplasmic variants with depth < 200x and low level heteroplasmies (HFs<5%) with depth < 500x, unless a Bonferroni-corrected p-value < 10^{-5} was given by deepSNV (*10, 11*). Specifically, we used deepSNV to compare the nucleotide counts at a putative heteroplasmic variant site in a sample with a random set of 500 other unrelated samples using a beta-binomial model, to further reduce the false positive calls which were caused by the NUMTs and the sequencing error.

Validating heteroplasmic variants - duplicates

In order to validate the reliability of variant calling in this study, we assessed 22 samples sequenced twice independently. The concordance of homoplasmic variants between duplicates was 100%. We further assessed the consistency of heteroplasmic variant calls. After quality control, there were 28 heteroplasmic variants with the confidence interval of VAF between 1% and 98% detected in 20 duplicates (40 samples) using MtoolBox. Only one of these variants, with a HF of 2.1%, was present in only one of the duplicates, giving an overall concordance of 96.4%. The deepSNV-based filter removed 6 non-concordant heteroplasmic calls with HF <3% (**fig. S2**). The implications of the duplicate analysis suggested the high accuracy of our analysis pipeline to detect the low level heteroplasmies.

Classifying the heteroplasmic variants

We defined three categories of heteroplasmy: (1) transmitted/inherited, if the variant was present in the mother and the offspring and was heteroplasmic in at least one of the two; (2) lost, if the heteroplasmic variant was present in the mother but not detected in the offspring; and (3) *de novo*, if the heteroplasmic variant was present in the offspring and not detected in the mother (**table S1 and Data S1**). Note that lost variants can be present in maternal blood but absent from the germ line. For the lost and *de novo* heteroplasmies, they are likely present in offspring (lost heteroplasmies) or mothers (*de novo* heteroplasmies) with very low level of heteroplasmic fraction (HF < 1%) which are below our detection threshold.

We studied the log2 ratio of HF between offspring and mothers after imputation of HF values below 1% to our detection threshold of 1%. Heteroplasmic variants with a positive/negative HS were deemed to have an increasing/decreasing HS over one generation. We further classified the heteroplasmic variants into two categories: (1) increasing HS, where the log2 ratio of HF between offspring and mothers > 0; (2) decreasing HS: where the log2 ratio of HF between offspring and mothers < 0.

Estimation of pairwise relatedness and nuclear ancestry

We estimated the degree of relatedness between individuals and categorised each individual's nuclear ancestry into one of European, African, Asian or Other using genotype data from a set of SNPs selected as follows. We identified the 292,878 autosomal SNPs typed by three widely used Illumina genotyping arrays (HumanCoreExome-12v1.1, HumanCoreExome-24v1.0 and HumanOmni2.5-8v1.1). Presence of a SNP in all three arrays indicates that it can be reliably measured by microarray and, hence, that it is likely to be reliably called using whole genome sequencing as well. To further ensure this, we removed SNPs with a missing genotype in at least one individual or with an overall pass rate below 0.99. We also removed SNPs at genomic positions in which more than two distinct alleles had been observed in the 1000 Genomes Phase 3 dataset (*50*) or the NIHR BioResource dataset to ensure that all genotypes could be coded unambiguously in as a count of the number of copies of the unique alternative allele carried by the individual. We then removed SNPs with a MAF < 0.3 in our dataset. Finally, we pruned SNPs using PLINK (v1.9) (http://pngu.mgh.harvard.edu/purcell/plink/) (*51*) so that all pairs of SNPs had an $r^2 < 0.2$. After filtering, 32,875 SNPs remained.

These well-measured, unlinked common SNPs were used to estimate relatedness in the non-admixed individuals within the 1000 Genomes Phase 3 data as follows. First, we ran the KING R package to compute an initial kinship matrix and identify a corresponding initial set of unrelated individuals (*52, 53*). We then used PC-AiR (*54, 55*) to perform a principal component analysis (PCA) on the standardised genotypes of these putatively unrelated individuals and project the standardised genotypes of the other individuals onto the fitted principal components (PCs). The PC-AiR object was then passed to the PC-Relate function to compute a kinship matrix that accounts for population structure as represented by the leading 20 PCs (*54*). Finally, the kinship matrix was passed to the PRIMUS (*56*) function to obtain a final set family relations and a final set of unrelated individuals based on kinship coefficients > 0.09.

We partitioned these unrelated individuals as non-Finnish Europeans, Finns, Africans, South Asians and East Asians using their 1000 Genomes population code annotations. Within each element of this partition, we modelled the score vectors of the leading five principal components by a multivariate normal distribution and estimated the corresponding mean vectors and covariance matrices. Subsequently, we projected the genotypes from each of the NIHR BioResource samples onto the vector space spanned by the leading five components of the 1000 Genomes PCA. We computed the likelihood of the projected data under the five multivariate normal models estimated from the 1000 Genomes scores and labelled the individual with the population corresponding to the model that yielded the highest likelihood, provided it was greater than 2000. Otherwise, the individual was labelled "Other." For the purposes of this work, we did not require distinguishing between non-Finnish Europeans and Finns because they share the same mitochondrial haplogroups. We therefore recorded only the greatest likelihood among the two populations and labelled it simply "European". For the same reason, we also recorded only the highest likelihood amongst East and South Asians and labelled it “Asian”. Finally, in order to exclude individuals with recent ancestry from more than one population (admixture), samples for which the model with the greatest likelihood was less than 10^{10} times greater than the model with the second greatest likelihood were labelled "Other".

To compute the relatedness within the new sequence dataset, the genotypes were merged with the 1000 Genomes genotypes and the procedure described above to compute relatedness was followed a second time. As the 1000 Genomes collection is genetically diverse by design, combining the two datasets ensured that a greater amount of genetic diversity was accounted for by the PCA than if only the new sequence dataset had been used. The PC-Relate estimates of the genome proportions shared by zero, one and two chromosomes between two individuals were used to categorise pairwise relations, identify mother-offspring pairs and the set of 10,210 unrelated individuals for the purposes of this work. The relatedness for all the mother and offspring pairs was further confirmed by the gender, age and the mitochondrial haplogroups. All the mother-offspring pairs shared the same mtDNA haplogroups.

Determining matched and mismatched groups

Mitochondrial DNA haplogroup assignment was performed using HaploGrep2 (*27, 57*). We compared the mtDNA phylogenetic haplogroup with the nuclear genetic ancestry in the same individual, and identified three distinct groups of individuals: (1) a haplogroup matched group where the mtDNA haplogroup was concordant with the nuclear ancestry; (2) a mismatched group where the nuclear ancestry and mtDNA were from different human populations, and (3) a final group where the nuclear ancestry could not be reliably determined (**Fig. 6, A and B and fig. S10**). In this particular analysis, we further excluded 42 out of the 12,975 samples because they had low haplogroup prediction scores as follows: 1) haplogroup prediction score below 0.8 in non-haplogroup H individuals, 2) haplogroup prediction score below 0.5 in haplogroup H and sub-haplogroup individuals. Because Haplogrep2 uses rCRS as the reference genome, which belongs to European haplogroup H (*49*), very few haplogroup specific variants were needed to predict the same haplogroup. Subsequent analyses focused on the haplogroup matched and mismatched groups.

Defining known and novel variants

Variants were considered to be novel if they were absent from 1000 Genomes datasets and dbSNP and were seen in at most one individual amongst 30,506 NCBI mtDNA sequences (*5*). The frequency of variants in 1000 Genomes and dbSNP was extracted using MToolBox (*8*). The allele frequencies in 30, 506 NCBI mtDNA sequences were computed in previous work (*5*).

Mutational spectra and signature

Mutational spectra were derived from the rCRS reference and alternative alleles at each variant site. The resulting spectra were composed of the relative frequencies of the six distinguishable point mutations (C:G>T:A, T:A>C:G, C:G>A:T, C:G>G:C, T:A>A:T and T:A>G:T). Each signature was displayed using a 96 substitution classification defined by the substitution class and the sequence context immediately 3' and 5' of the mutated base (*25, 58*).

Conservation scores

PhastCons100way scores for the mtDNA genome were downloaded from UCSC (*42*). We defined as conserved sites those with a PhastCons100way score >0.5, and the non-conserved sites as those with a PhastCons100way score <0.5. We estimated the number of expected conserved/non-conserved sites as the number of base pairs in the entire mtDNA genome with a PhaseCons 100 score above/below 0.5.

Probability of maternal mtDNA transmission

In order to understand the determinants of maternal transmission of heteroplasmic variants, we are modelling the probability of transmission of 1,043 heteroplasmic variants observed in the mothers using the following logistic regression model:

$$\text{logit}\, P\left(y_{ijl} = 1\right) = \alpha + \beta_1 \mathbf{1}_{j=1} + \beta_2 \mathbf{1}_{j=2} + \beta_3 \mathbf{1}_{j=3} + \beta_4 \mathbf{1}_{j=4} + \gamma w_{ijl} + \eta z_{ijl}$$

where $y_{ijl} = 1$ if the lth variant within mitochondrial genomic region j in mother i was transmitted and zero otherwise; $j = 0, 1, 2, 3$ or 4 denote the coding, Dloop, rRNA, tRNA and the remainder sequences, respectively; w_{ijl} is the logit of the HF of the lth variant within mitochondrial genomic region j in mother i; and $z_{ijl} = 1$ if the lth variant within mitochondrial genomic region j in mother i was observed in no individuals from the 1000 Genomes datasets, dbSNP and at most one individual amongst 30,506 NCBI mtDNAs, otherwise it was equal to zero. The model was fitted using the glm() function in R(v3.3) (http://CRAN.R-project.org/). The predictive accuracy was assessed using the mean area under the curve over repeated 10-fold cross-validations obtained using R.

Homoplasmic allele frequency in the population and heteroplasmic variants

To explore the relationship between the homoplasmic allele frequency in the general population and the rate at which individuals who are not homoplasmic for the alternate allele are heteroplasmic, we fitted a logistic regression model as follows. For each variant observed in 30,506 NCBI sequences and individual who was not homoplasmic for the alternate allele, we recorded whether the individual was heteroplasmic or not as a binary variable. We then fitted a logistic regression model with an intercept and a single continuous covariate encoding the logit of the allele frequency in 30,506 NCBI sequences.

Defining haplogroup specific variants on mtDNA phylogenetic tree

We extracted 4,476 SNVs present on mtDNA phylogenetic tree (*27*) using in-house python scripts. We then focused on SNVs either present in only one super-population (European, Asian or African) on the phylogenetic tree, or present on two super-population, but were commonly seen in one population (>1%) and not seen or extremely rare in the other population in 17,520 full mtDNA sequences from NCBI database (*5*). We defined 2,641 haplogroup-specific variants, including 426 African variants, 1,275 Asian variants and 940 European variants.

Nuclear genome ancestry and mtDNA heteroplasmic variants

In order to understand the determinants of heteroplasmic variation, we modelled the presence or absence of a heteroplasmic variant in a particular individual using logistic regression. We considered only the 2,215 mtDNA variants from of over 4,000 haplogroup-specific variants that are present exclusively in European or Asian branches of the world mtDNA phylogeny (*27*), as this allows unambiguous assignation of mitochondrial ancestry to each variant. To avoid the potential for bias induced by recent shared ancestry between individuals, we considered only the 9,631 unrelated individuals in matched and mismatched groups. We fitted the following logistic regression model:

$$\text{logit}\, P\left(y_{ij} = 1\right) = \alpha + \beta_1 \mathbf{1}_{\boldsymbol{x_j=1}} + \beta_2 \mathbf{1}_{\boldsymbol{x_j=2}} + \ \gamma w_j + \eta \mathbf{1}_{\boldsymbol{z_i=w_i}} + \omega \mathbf{1}_{\boldsymbol{x_j=w_i \cap z_i=w_i}} + \ \psi \mathbf{1}_{\boldsymbol{x_j=w_i \cap z_i \neq w_i}}$$

where $y_{ij} = 1$ if variant j is heteroplasmic in individual i, and zero otherwise; $x_j = 0, 1$ or 2 depending on whether the variant ancestry is Asian, African or European, respectively; w_j is the logit of the homoplasmic allele frequency of variant j in 30,506 NCBI samples; $z_i = 0, 1$ or 2 depending on whether the mitochondrial ancestry of individual i is Asian, African or European, respectively; and $w_i = 0, 1$ or 2 depending on whether the nuclear ancestry of individual i is Asian, African or European, respectively. The indicator variable **1** evaluates to 1 if the conditions in its subscript are met and zero otherwise. The model was fitted using the glm() function in R.

Validation dataset

We repeated the nuclear-mtDNA ancestry analysis in 42,799 WGS from whole-blood samples from the Genomics England 100,000 Genomes Rare Disease Main Programme aligned to GRCh37 or/and hg38. The same bioinformatic pipeline was used to call mtDNA variants as in the observation dataset (see above). 40,325 mtDNA sequences had mtDNA coverage > 99.9%, mean depth > 500x, no more than 20 heteroplasmies and an accurate prediction of mtDNA ancestry (see above) and were thus retained for analysis. The mean depth of WGS was 42x (range 29x to 121x) and the mean depth of mtDNA was 1919x (range 500x to 7458x) (**fig S15**).

To infer ethnicity and relatedness in this dataset, we lifted over our original list of common SNPs from GRCh37 to GRCh38 using the liftOver tool from the UCSC web browser (https://genome.ucsc.edu/cgi-bin/hgLiftOver). During this conversion 210 SNPs were lost: 1) 17 SNPs because their positions could not be mapped to GRCh38, 2) 134 SNPs because the reference and alternate alleles had been swapped, 3) 42 SNPs because they were missing from the GRCh38 1000G VCF, 4) 17 because their AFs in Europeans were discrepant between the observation and the validation datasets. We then ran the same code as described above but using the list of 32,665 SNPs on GRCh38.

We implemented a custom heuristic procedure to identify the maximal unrelated set, instead of using PRIMUS, as PRIMUS was computationally unable to cope with such a large dataset. The procedure iteratively removes the individual with the greatest number of relatives (defined as individuals with a kinship coefficient >0.09) until there are no relatives remaining. To boost power, we maximised the number of mismatched individuals in our unrelated set as follows.

First, we identified the maximum unrelated set amongst mismatched individuals. We then combined this set with all matched individuals and re-ran the heuristic approach to obtain a final set of unrelated individuals.

Statistical analysis

All statistical analyses in this study were suggested in the text and performed using R (v3.3) (http://CRAN.R-project.org/). Circos plots were made using Circos (*59*). Figures were generated using Matplotlib (https://matplotlib.org) in Python (http://www.python.org) and R.

Collaborators

NIHR BioResource - Rare Disease and Control Collections: NIHR BioResource Core Teams Anthony Attwood[3,4], Matthew Brown[3,4], Naomi Clements Brod[3,4], Abigail Crisp-Hihn[3,4], John Davis[3,4], Sri V V Deevi[3,4], Eleanor F Dewhurst[3,4], Karen Edwards[3,4], Marie Erwood[3,4], James Fox[3,4], Amy J Frary[3,4], Fengyuan Hu[3,4], Jennifer Jolley[3,4], Nathalie Kingston[4], Rachel Linger[4,43], Rutendo Mapeta[3,4], Jennifer Martin[4,22,43], Stuart Meacham[3,4], Sofia Papadia[4,43], Paula J Rayner-Matthews[3,4], Crina Samarghitean[3,4], Olga Shamardina[3,4], Ilenia Simeoni[3,4], Simon Staines[3,4], Emily Staples[22], Hannah Stark[4,43], Jonathan Stephens[3,4], Catherine Titterton[3,4], Salih Tuna[3,4], Julie von Ziegenweidt[3,4], Christopher Watt[3,4], Deborah Whitehorn[3,4], Yvette Wood[3,4], Katherine Yates[3,4,22], Ping Yu[3,4], Roger James[3,4], Sofie Ashford[4,43], Christopher J Penkett[3,4], Kathleen E Stirrups[3,4]; **Bleeding, thrombotic and Platelet Disorders (BPD)** Tadbir Bariana[12,50], Claire Lentaigne[51,52], Suthesh Sivapalaratnam[45,53,54,55], Sarah K Westbury[56,57], David J Allsup[58,59], Tamam Bakchoul[60], Tina Biss[61], Sara Boyce[62], Janine Collins[3,53], Peter W Collins[63], Nicola S Curry[64], Kate Downes[3,4], Tina Dutt[65], Wendy N Erber[66], Gillian Evans[67], Tamara Everington[68,69], Remi Favier[70,71], Keith Gomez[12,50], Daniel Greene[3,4,49], Paolo Gresele[72], Daniel Hart[53], Rashid Kazmi[62], Anne M Kelly[24], Michele Lambert[73,74], Bella Madan[75], Sarah Mangles[69], Mary Mathias[76], Carolyn Millar[51,52], Samya Obaji[77], Kathelijne Peerlinck[78], Catherine Roughley[67], Sol Schulman[79], Marie Scully[80], Susan E Shapiro[64], Keith Sibson[76], Ilenia Simeoni[3,4], Matthew C Sims[3,81], R Campbell Tait[82], Kate Talks[61], Chantal Thys[78], Cheng-Hock Toh[65], Chris Van Geet[78], John-Paul Westwood[80], Sofia Papadia[4,43], Ernest Turro[3,4,49], Andrew D Mumford[56,57], Willem H Ouwehand[3,4,44,45,46], Kathleen Freson[78], Michael A Laffan[51,52]; **Cerebral Small Vessel Disease (CSVD)** Rhea Y Y Tan[30], Kirsty Harkness[83], Sarju Mehta[84], Keith W Muir[85], Ahamad Hassan[86], Matthew Traylor[30], Anna M Drazyk[30], Hugh S Markus[30]; **Ehlers Danlos Syndrome (EDS)** David Parry[7], Munaza Ahmed[87], Hanadi Kazkaz[80], Anthony M Vandersteen[88], Timothy J Aitman[6,7]; **Hypertrophic Cardiomyopathy (HCM)** Elizabeth Ormondroyd[36,37], Kate Thomson[36,37], Timothy Dent[37], Paul Brennan[15,16,17], Rachel J Buchan[89,90], Teofila Bueser[21,26,91], Gerald Carr-White[20], Stuart Cook[89,92,93,94], Matthew J Daniels[36,37,95], Andrew R Harper[35,36,89], James S Ware[89,90,92], Hugh Watkins[35,36,37]; **Intrahepatic Cholestasis of Pregnancy (ICP)** Peter H Dixon[41], Jenny Chambers[41,96], Floria Cheng[96], Maria C Estiu[97], William M Hague[98], Hanns-Ulrich Marschall[99], Marta Vazquez-Lopez[96], Catherine Williamson[41,42]; **Inherited Retinal Disorders (IRD)** Gavin Arno[38,39], Eleanor F Dewhurst[3,4], Marie Erwood[3,4], Courtney E French[100], Michel Michaelides[38], Anthony T Moore[38,39,101], Alba Sanchis-Juan[3,4], Keren Carss[3,4], Andrew R Webster[38,39], F Lucy Raymond[4,28]; **Leber Hereditary Optic Neuropathy (LHON)** Patrick F Chinnery[1,2,4], Philip Griffiths[102,103], Rita Horvath[34,104], Gavin Hudson[34], Neringa Jurkute[12,38], Angela Pyle[34], Wei Wei[1,2], Patrick Yu-Wai-Man[1,2,38,39]; **Multiple Primary Malignant Tumours (MPMT)** James Whitworth[28,29,105], Julian Adlard[106], Munaza Ahmed[87], Ruth Armstrong[28,29,105], Carole Brewer[107], Ruth Casey[28,29,105], Trevor R P Cole[108], Dafydd Gareth Evans[109], Lynn Greenhalgh[110], Helen L Hanson[111], Jonathan Hoffman[108], Louise Izatt[112], Ajith Kumar[87], Fiona Lalloo[14], Kai Ren Ong[108], Soo-Mi Park[29,105,113], Claire Searle[114], Lucy Side[87], Katie Snape[111], Emma Woodward[14], Marc Tischkowitz[113,115], Eamonn R Maher[28,29]; **Neurological and Developmental Disorders (NDD)** Keren Carss[3,4], Eleanor F Dewhurst[3,4], Marie Erwood[3,4], Courtney E French[100], Detelina Grozeva[28], Alba Sanchis-Juan[3,4], Manju A Kurian[116,117], F Lucy Raymond[4,28]; **Neuropathic Pain Disorders (NPD)** Andreas C Themistocleous[10], David Gosal[118], Rita Horvath[34,104], Andrew Marshall[119,120,121], Emma Matthews[122,123], Mark I McCarthy[33,35,124], Tara Renton[91], Andrew S C Rice[125,126], Tom Vale[10],

Suellen M Walker[9,127], Christopher Geoffrey Woods[24,28], David L Bennett[10]; **Primary Immune Disorders (PID)** James E Thaventhiran[128], Hana Lango Allen[3,4], Sinisa Savic[129,130], Hana Alachkar[118], Richard Antrobus[131], Helen E Baxendale[22,23,100,132], Michael J Browning[133], Matthew S Buckland[134], Nichola Cooper[135], J David M Edgar[136,137], William Egner[138], Kimberly C Gilmour[9,127], Sarah Goddard[139], Pavels Gordins[140], Sofia Grigoriadou[141], Scott Hackett[142], Rosie Hague[143], Grant Hayman[144], Archana Herwadkar[118], Aarnoud P Huissoon[142], Stephen Jolles[145], Peter Kelleher[146,147], Dinakantha Kumararatne[148], Hilary Longhurst[141], Lorena E Lorenzo[141], Paul A Lyons[22], Jesmeen Maimaris[127], Sadia Noorani[149], Alex Richter[131], Crina Samarghitean[3,4], Ravishankar B Sargur[138], W A Carrock Sewell[150], Ilenia Simeoni[3,4], Emily Staples[22], David Thomas[22], Moira J Thomas[151,152], Austen Worth[9], Patrick F K Yong[153], Taco W Kuijpers[154,155], Adrian J Thrasher[127], Kenneth G C Smith[22]; **Primary Membranoproliferative Glomerulonephritis (PMG)** Adam P Levine[11], Omid Sadeghi-Alavijeh[11], Edwin K S Wong[16,156], H Terence Cook[157], Melanie M Y Chan[11], Matthew Hall[158], Claire Harris[156], Paul McAlinden[156], Kevin J Marchbank[156,159], Stephen Marks[9], Heather Maxwell[143], Monika Mozere[11], Julie Wessels[139], Sally A Johnson[156,160], Daniel P Gale[11]; **Pulmonary Arterial Hypertension (PAH)** Marta Bleda[22], Charaka Hadinnapola[22], Matthias Haimel[3,4,22], Emilia Swietlik[22], Harm Bogaard[161], Colin Church[162], Gerry Coghlan[134], Robin Condliffe[163], Paul Corris[156,164], Cesare Danesino[165], Mélanie Eyries[166], Henning Gall[167], Hossein-Ardeschir Ghofrani[135,167], J Simon R Gibbs[89], Barbara Girerd[168,169,170], Simon Holden[84], Arjan Houweling[161], Luke S Howard[89,171], Marc Humbert[168,169,170], David G Kiely[163], Gabor Kovacs[172,173], Allan Lawrie[174], Robert V MacKenzie Ross[175], Jennifer Martin[4,22,43], Shahin Moledina[9], David Montani[168,169,170], Michael Newnham[22,23], Andrea Olschewski[172], Horst Olschewski[172,173], Andrew Peacock[162], Joanna Pepke-Zaba[23], Laura Scelsi[176], Werner Seeger[167], Florent Soubrier[166], Jay Suntharalingam[175], Mark Toshner[22,23], Carmen Treacy[22,23], Richard Trembath[26], Anton Vonk Noordegraaf[161], Quinten Waisfisz[177], John Wharton[135], Martin R Wilkins[135], Stephen J Wort[90,178], Katherine Yates[3,4,22], Stefan Gräf[3,4,22], Nicholas W Morrell[4,22]; **Stem cell and Myeloid Disorders (SMD)** Eleni Louka[31,33], Noemi B Roy[31,32,33], Anupama Rao[9], Philip Ancliff[9], Christian Babbs[31,33], D Mark Layton[51,52], Adam J Mead[31], Jennifer O'Sullivan[75], Steven Okoli[31,33], Irene Roberts[31,32,33]; **Steroid Resistant Nephrotic Syndrome (SRNS)** Moin Saleem[179,180], Agnieszka Bierzynska[179], Carmen Bugarin Diz[26], Elizabeth Colby[179], Melanie N Ekani[20], Simon Satchell[179,181], Ania Koziell[26,27]

100,000 Genomes Project Rare Diseases Pilot Sites: Genomics England Core Teams Tom Fowler[5], Augusto Rendon[3,5], Richard Scott[5,9], Damian Smedley[5,48], Katherine R Smith[5], Ellen Thomas[5,20], Mark Caulfield[5,48]; **Cambridge University Hospitals NHS Foundation Trust** Stephen Abbs[113], Nigel Burrows[24], Manali Chitre[100], Eleanor F Dewhurst[3,4], R Andres Floto[22,23,24], Michael Gattens[24], Mark Gurnell[22,24], Simon Holden[84], Wilf Kelsall[24], Sarju Mehta[84], Ken E S Poole[22,24], Robert Ross-Russell[24], Olivera Spasic-Boskovic[113], Philip Twiss[113], Annette Wagner[24], F Lucy Raymond[4,28]; ,**Central Manchester University Hospitals NHS Trust and Manchester University** Siddharth Banka[13,14], Graeme C Black[13,14], Jill Clayton-Smith[13,14], Sofia Douzgou[13,14], William G Newman[13,14]; **Great Ormond Street Hospital for Children NHS Foundation Trust and University College London** Lara Abulhoul[9], Paul Aurora[9], Detlef Bockenhauer[9], Maureen Cleary[9], Mehul Dattani[8,182], Vijeya Ganesan[9], Clarissa Pilkington[9], Shamima Rahman[8,9], Neil Shah[9,127], Lucy Wedderburn[127,183,184], Maria A K Bitner-Glindzicz[8,9]; **Guys and St Thomas' Hospital NHS Foundation Trust and King's College London** Teofila Bueser[21,185], Cecilia J Compton[21], Charu Deshpande[21], Hiva Fassihi[186], Eshika

Haque[21], Louise Izatt[21], Dragana Josifova[21], Shehla N Mohammed[21], Leema Robert[21], Sarah J Rose[21], Deborah M Ruddy[21], Robert N Sarkany[186], Genevieve Sayer[21], Adam C Shaw[21], Melita Irving[21], Frances A Flinter[21]; **Oxford University Hospitals NHS Trust and the University of Oxford** Carolyn Campbell[187], Kate Gibson[187], Nils Koelling[40], Tracy Lester[187], Andrea H Nemeth[10,188], Claire Palles[189], Smita Patel[190], Noemi B Roy[40,191], Arjune Sen[33,192,193], John Taylor[187], Ian P Tomlinson[189], Jenny C Taylor[33,35], Andrew O M Wilkie[33,37,40]; **Moorfields Eye Hospital NHS Trust and University College London** Gavin Arno[38,39], Samantha Malka[38,39], Michel Michaelides[38,39], Anthony T Moore[38,39,101], Andrew R Webster[38,39]; **Newcastle upon Tyne Hospitals NHS Foundation Trust and Newcastle University** Paul Brennan[17], Andrew C Browning[194], John Burn[17], Patrick F Chinnery[1,2,4], Anthony De Soyza[16,156,195], Jodie Graham[196], Rita Horvath[102], Simon Pearce[16,196], Richard Quinton[16,34], Andrew M Schaefer[16,102], Brian T Wilson[16,34,87], Michael Wright[17], Patrick Yu-Wai-Man[1,2,197], John A Sayer[16,34]; **University College London Hospitals NHS Trust and University College London** Michael Simpson[198], Petros Syrris[18], Perry Elliott[18,19], Henry Houlden[25], Phil L Beales[8,9]

Author addresses and affiliations

[1]Department of Clinical Neurosciences, School of Clinical Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [2]Medical Research Council Mitochondrial Biology Unit, Cambridge Biomedical Campus, Cambridge, UK. [3]Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [4]NIHR BioResource, Cambridge University Hospitals NHS Foundation, Cambridge Biomedical Campus, Cambridge, UK. [5]Genomics England, Charterhouse Square, London, UK. [6]MRC Clinical Sciences Centre, Faculty of Medicine, Imperial College London, London, UK. [7]Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [8]Genetics and Genomic Medicine Programme, UCL Great Ormond Street Institute of Child Health, London, UK. [9]Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. [10]The Nuffield Department of Clinical Neurosciences, University of Oxford, John Radcliffe Hospital, Oxford, UK. [11]UCL Centre for Nephrology, University College London, London, UK. [12]University College London, London, UK. [13]Evolution and Genomic Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK. [14]Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester Universities Foundation NHS Trust, Manchester, UK. [15]Newcastle University, Newcastle upon Tyne, UK. [16]Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [17]Northern Genetics Service, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [18]UCL Institute of Cardiovascular Science, University College London, London, UK. [19]Barts Heart Centre, St Bartholomew's Hospital, Barts Health NHS Trust, London, UK. [20]Guy's and St Thomas' Hospital, Guy's and St Thomas' NHS Foundation Trust, London, UK. [21]Clinical Genetics Department, Guy's and St Thomas NHS Foundation Trust, London, UK. [22]Department of Medicine, School of Clinical Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [23]Royal Papworth Hospital NHS Foundation Trust, Cambridge, UK. [24]Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [25]Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK. [26]King's College London, London, UK. [27]Department of Paediatric Nephrology, Evelina London Children's Hospital, Guy's & St Thomas' NHS Foundation Trust, London, UK. [28]Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [29]NIHR Cambridge Biomedical Research

Centre, Cambridge Biomedical Campus, Cambridge, UK. [30]Stroke Research Group, Department of Clinical Neurosciences, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [31]MRC Molecular Haematology Unit, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [32]Department of Paediatrics, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [33]NIHR Oxford Biomedical Research Centre, Oxford University Hospitals Trust, Oxford, UK. [34]Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK. [35]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [36]Department of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [37]Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [38]Moorfields Eye Hospital NHS Foundation Trust, London, UK. [39]UCL Institute of Ophthalmology, University College London, London, UK. [40]MRC Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK. [41]Division of Women's Health, King's College London, London, UK. [42]Institute of Reproductive and Developmental Biology, Surgery and Cancer, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. [43]Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [44]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [45]NHS Blood and Transplant, Cambridge Biomedical Campus, Cambridge, UK. [46]British Heart Foundation Cambridge Centre of Excellence, University of Cambridge, Cambridge, UK. [47]Department of Renal Medicine, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [48]William Harvey Research Institute, NIHR Biomedical Research Centre at Barts, Queen Mary University of London, London, UK. [49]MRC Biostatistics Unit, Cambridge Institute of Public Health, University of Cambridge, Cambridge, UK. [50]The Katharine Dormandy Haemophilia Centre and Thrombosis Unit, Royal Free London NHS Foundation Trust, London, UK. [51]Department of Haematology, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. [52]Centre for Haematology, Imperial College London, London, UK. [53]The Royal London Hospital, Barts Health NHS Foundation Trust, London, UK. [54]Department of Haematology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [55]Queen Mary University of London, London, UK. [56]School of Cellular and Molecular Medicine, University of Bristol, Bristol, UK. [57]University Hospitals Bristol NHS Foundation Trust, Bristol, UK. [58]Queens Centre for Haematology and Oncology, Castle Hill Hospital, Hull and East Yorkshire NHS Trust, Cottingham, UK. [59]Hull York Medical School, University of Hull, Hull, UK. [60]Center for Clinical Transfusion Medicine, University Hospital of Tübingen, Tübingen, Germany. [61]Haematology Department, Royal Victoria Infirmary, The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [62]Southampton General Hospital, University Hospital Southampton NHS Foundation Trust, Southampton, UK. [63]Institute of Infection and Immunity, School of Medicine Cardiff University, Cardiff, UK. [64]Oxford Haemophilia and Thrombosis Centre, Oxford University Hospitals NHS Trust, Oxford Comprehensive Biomedical Research Centre, Oxford, UK. [65]The Roald Dahl Haemostasis and Thrombosis Centre, The Royal Liverpool University Hospital, Liverpool, UK. [66]Medical School and School of Biomedical Sciences, Faculty of Health and Medical Sciences, The University of Western Australia, Crawley, Australia. [67]Haemophilia Centre, Kent & Canterbury Hospital, East Kent Hospitals University Foundation Trust, Canterbury, UK. [68]Salisbury District Hospital, Salisbury NHS Foundation Trust, Salisbury, UK. [69]Haemophilia, Haemostasis and Thrombosis Centre, Hampshire Hospitals NHS Foundation Trust, Basingstoke, UK. [70]INSERM UMR 1170, Gustave Roussy Cancer Campus, Universite Paris-Saclay, Villejuif, France. [71]Service d'Hematologie biologique, Centre de Reference des Pathologies Plaquettaires, Hopital Armand Trousseau, Assistance Publique-Hopitaux de Paris, Paris, France. [72]Section of Internal and Cardiovascular Medicine, University of Perugia, Perugia, Italy. [73]Division of Hematology, The Children's Hospital of Philadelphia, Philadelphia, USA. [74]Department of Pediatrics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, USA. [75]Department of Haematology, Guy's and St Thomas' NHS Foundation Trust, London, UK. [76]Department of Haematology, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. [77]The Arthur Bloom Haemophilia Centre, University Hospital of Wales, Cardiff, UK. [78]Department of Cardiovascular Sciences, Center for Molecular and Vascular Biology, KULeuven, Leuven, Belgium. [79]Beth Israel Deaconess Medical Centre and Harvard Medical School, Boston, USA. [80]University College London Hospitals NHS Foundation Trust, London, UK. [81]Oxford Haemophilia and Thrombosis Centre, The Churchill Hospital, Oxford University Hospitals NHS Trust, Oxford, UK. [82]Glasgow Royal Infirmary, NHS Greater Glasgow and Clyde, Glasgow, UK. [83]Department of Neurology, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK. [84]Department of Clinical Genetics, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [85]Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK. [86]Department of Neurology, Leeds Teaching Hospital NHS Trust, Leeds, UK. [87]North East Thames Regional Genetics Service, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. [88]Division of Medical Genetics, IWK Health Centre, Dalhousie University, Halifax, Canada. [89]National Heart and Lung Institute, Imperial College London, London, UK. [90]Royal Brompton Hospital, Royal Brompton and Harefield NHS Foundation Trust, London, UK. [91]King's College Hospital NHS Foundation Trust, London, UK. [92]MRC London Institute of Medical Sciences, Imperial College London, London, UK. [93]National Heart Research Institute Singapore, National Heart Centre Singapore, Singapore, Singapore. [94]Division of Cardiovascular & Metabolic Disorders, Duke-National University of Singapore, Singapore, Singapore. [95]Department of Biotechnology, Graduate School of Engineering, Osaka University, Suita, Osaka, Japan. [96]Women’s Health Research Centre, Surgery and Cancer, Faculty of Medicine, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. [97]Ramón Sardá Mother’s and Children’s Hospital, Buenos Aires, Argentina. [98]Robinson Research Institute, Discipline of Obstetrics and Gynaecology, The University of Adelaide, Women's and Children's Hospital, Adelaide, Australia. [99]Department of Molecular and Clinical Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. [100]Department of Paediatrics, School of Clinical Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. [101]Ophthalmology Department, UCSF School of Medicine, San Francisco, USA. [102]Wellcome Centre for Mitochondrial Research, Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK. [103]Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK. [104]John Walton Muscular Dystrophy Research Centre, Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK. [105]Cancer Research UK Cambridge Centre, Cambridge Biomedical Campus, Cambridge, UK. [106]Yorkshire Regional Genetics Service, Chapel Allerton Hospital, Leeds Teaching Hospitals NHS Trust, Leeds, UK. [107]Department of Clinical Genetics, Royal Devon & Exeter Hospital, Royal Devon and Exeter NHS Foundation Trust, Exeter, UK. [108]West Midlands Regional Genetics Service, Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK. [109]Manchester University NHS Foundation Trust, Manchester, UK. [110]Department of Clinical Genetics, Liverpool Women's NHS Foundation, Liverpool, UK. [111]Department of Clinical Genetics, St George's

University Hospitals NHS Foundation Trust, London, UK. [112]Department of Clinical Genetics, Guy's and St Thomas' NHS Foundation Trust, London, UK. [113]East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [114]Department of Clinical Genetics, Nottingham University Hospitals NHS Trust, Nottingham, UK. [115]Department of Medical Genetics and NIHR Cambridge Biomedical Research Centre, University of Cambridge, Cambridge, UK. [116]Developmental Neurosciences, UCL Great Ormond Street Institute of Child Health, London, UK. [117]Department of Neurology, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK. [118]Salford Royal NHS Foundation Trust, Salford, UK. [119]Faculty of Medical and Human Sciences, Centre for Endocrinology and Diabetes, Institute of Human Development, University of Manchester, Manchester, UK. [120]Department of Clinical Neurophysiology, Manchester Royal Infirmary, Central Manchester University Hospitals National Health Service Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK. [121]National Institute for Health Research/Wellcome Trust Clinical Research Facility, Manchester, UK. [122]The National Hospital for Neurology and Neurosurgery, University College London Hospitals NHS Foundation Trust, London, UK. [123]MRC Centre for Neuromuscular Diseases, Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK. [124]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Oxford, UK. [125]Pain Research, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London, UK. [126]Pain Medicine, Chelsea and Westminster Hospital NHS Foundation Trust, London, UK. [127]UCL Great Ormond Street Institute of Child Health, London, UK. [128]MRC Toxicology Unit, School of Biological Sciences, University of Cambridge, Cambridge, UK. [129]The Department of Clinical Immunology and Allergy and The NIHR Leeds Biomedical Research Centre, Leeds, UK. [130]Leeds Institute of Rheumatic and Musculoskeletal Medicine, St James’s University Hospital, Leeds, UK. [131]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [132]Division of Clinical Biochemistry and Immunology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [133]Department of Immunology, Leicester Royal Infirmary, Leicester, UK. [134]Royal Free London NHS Foundation Trust, London, UK. [135]Department of Medicine, Imperial College London, London, UK. [136]Regional Immunology Service, The Royal Hospitals, Belfast, UK. [137]Queen's University Belfast, Belfast, UK. [138]Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK. [139]University Hospitals of North Midlands NHS Trust, Stoke-on-Trent, UK. [140]East Yorkshire Regional Adult Immunology and Allergy Unit, Hull Royal Infirmary, Hull & East Yorkshire Hospitals NHS Trust, Hull, UK. [141]Barts Health NHS Foundation Trust, London, UK. [142]Birmingham Heartlands Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [143]Royal Hospital for Children, NHS Greater Glasgow and Clyde, Glasgow, UK. [144]Epsom & St Helier University Hospitals NHS Trust, London, UK. [145]Immunodeficiency Centre for Wales, University Hospital of Wales, Cardiff, IUK. [146]Centre for Immunology & Vaccinology , Chelsea & Westminster Hospital , Department of Medicine, Imperial College London, London, UK. [147]Department of Respiratory Medicine Royal Brompton & Harefield NHS Foundation Trust, London, UK. [148]Department of Clinical Immunology, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [149]Sandwell and West Birmingham Hospitals NHS Trust, Birmingham, UK. [150]Scunthorpe General Hospital, Northern Lincolnshire and Goole NHS Foundation Trust, Scunthorpe, UK. [151]Gartnavel General Hospital, NHS Greater Glasgow and Clyde, Glasgow, UK. [152]Queen Elizabeth University Hospital, Glasgow, UK. [153]Frimley Park Hospital, NHS Frimley Health Foundation Trust, Camberley, UK.

[154]Department of Pediatric Hematology, Immunology, Rheumatology and Infectious Diseases, Emma Children’s Hospital, Academic Medical Center (AMC), University of Amsterdam, Amsterdam, The Netherlands. [155]Department of Blood Cell Research, Sanquin, Amsterdam, The Netherlands. [156]Institute of Cellular Medicine, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK. [157]Imperial College Renal and Transplant Centre, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK. [158]Nottingham University Hospitals NHS Trust, Nottingham, UK. [159]The National Renal Complement Therapeutics Centre, Royal Victoria Infirmary, Newcastle upon Tyne, UK. [160]Department of Paediatric Nephrology, Great North Children’s Hospital, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [161]Department of Pulmonary Medicine, VU University Medical Centre, Amsterdam, The Netherlands. [162]Golden Jubilee National Hospital, Glasgow, UK. [163]Sheffield Pulmonary Vascular Disease Unit, Royal Hallamshire Hospital NHS Foundation Trust, Sheffield, UK. [164]National Pulmonary Hypertension Service (Newcastle), The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [165]Department of Molecular Medicine, General Biology, and Medical Genetics Unit, University of Pavia, Pavia, Italy. [166]Departement de Genetique & ICAN, Hopital Pitie-Salpetriere, Assistance Publique Hopitaux de Paris, Paris, France. [167]University of Giessen and Marburg Lung Center (UGMLC), Giessen, Germany. [168]Univ. Paris-Sud, Faculty of Medicine, University Paris-Saclay, Le Kremlin Bicetre, France. [169]Service de Pneumologie, Centre de Reference de l'Hypertension Pulmonaire, Hopital Bicetre (Assistance Publique Hopitaux de Paris), Le Kremlin Bicetre, France. [170]INSERM U999, Hospital Marie Lannelongue, Le Plessis Robinson, France. [171]National Pulmonary Hypertension Service, Imperial College Healthcare NHS Trust, London, UK. [172]Ludwig Boltzmann Institute for Lung Vascular Research, Graz, Austria. [173]Dept of Internal Medicine, Division of Pulmonology, Medical University of Graz, Graz, Austria. [174]Department of Infection, Immunity & Cardiovascular Disease, University of Sheffield, Sheffield, UK. [175]Royal United Hospitals Bath NHS Foundation Trust, Bath, UK. [176]Division of Cardiology, Fondazione IRCCS Policlinico S. Matteo, Pavia, Italy. [177]Department of Clinical Genetics, VU University Medical Centre, Amsterdam, The Netherlands. [178]Imperial College London, London, UK. [179]Bristol Renal and Children's Renal Unit, Bristol Medical School, University of Bristol, Bristol, UK. [180]Bristol Royal Hospital for Children, University Hospitals Bristol NHS Foundation Trust, Bristol, UK. [181]North Bristol NHS Trust, Bristol, UK. [182]London Centre for Paediatric Endocrinology and Diabetes, Great Ormond Street Hospital for Children, London, UK. [183]NIHR Great Ormond Street Biomedical Research Centre, London, UK. [184]Arthritis Research UK Centre for Adolescent Rheumatology at UCL UCLH and GOSH, London, UK. [185]Florence Nightingale Faculty of Nursing, Midwifery & Palliative Care, King's College London, London, UK. [186]St Johns Institute of Dermatology, St Thomas' Hospital, Guy's and St Thomas' NHS Foundation Trust, London, UK. [187]Oxford Medical Genetics Laboratories, Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [188]Department of Clinical Genetics, Churchill Hospital, Oxford University Hospitals NHS Trust, Oxford, UK. [189]Institute of Cancer and Genomic Sciences, Institute of Biomedical Research, University of Birmingham, Birmingham, UK. [190]Department of Clinical Immunology, John Radcliffe Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [191]Department of Haematology, Oxford University Hospital Foundation Trust, Oxford, UK. [192]Oxford Epilepsy Research Group, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK. [193]Nuffield Department of Surgery, University of Oxford, Oxford, UK. [194]Newcastle Eye Centre, Royal Victoria Infirmary, The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon

Tyne, UK. [195]NIHR Centre for Aging, Newcastle University, Newcastle upon Tyne, UK. [196]Newcastle BRC, Newcastle University, Newcastle upon Tyne, UK. [197]NIHR Biomedical Research Centre at Moorfields Eye Hospital and UCL Institute of Ophthalmology, London, UK. [198]Genetics and Molecular Medicine, King's College London, London, UK.

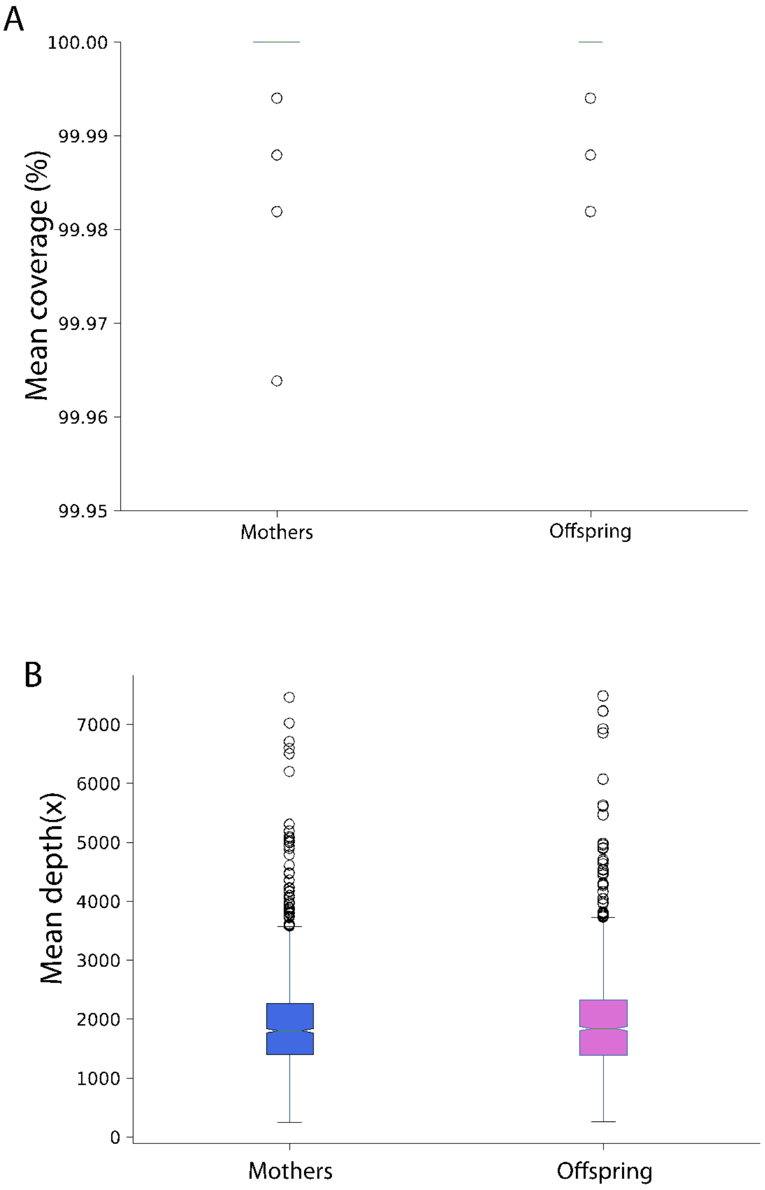


Fig. S1. Sequencing coverage and depth. (**A**) Box-whisker plots of mean coverage of mtDNA genome in 1,526 mother-offspring pairs. (**B**) Box-whisker plots of mean read depth of mtDNA genome in 1,526 mother-offspring pairs. There was no significant difference between the mtDNA sequence depth of mother and offspring (*P*=0.49, two sample *t*-test).

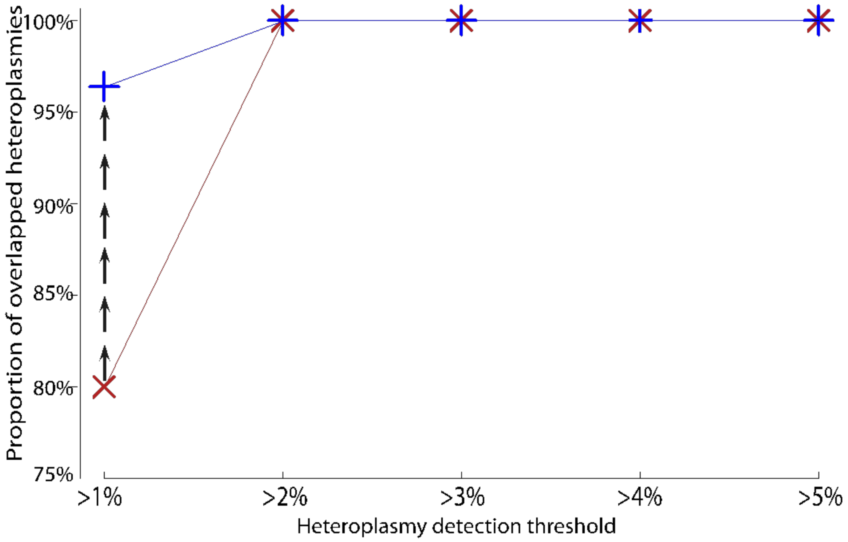


Fig. S2. Reproducibility of the heteroplasmy calling pipeline. We assessed 22 samples sequenced independently on two occasions. After quality control, there were 28 heteroplasmic variants with the confidence interval of VAF between 1% and 98% detected in 20 duplicates (40 samples) using MtoolBox. Only one of these variants, with a HF of 2.1%, was present in only one of the duplicates, giving an overall concordance of 96.4%. The deepSNV-based filter removed 6 non-concordant heteroplasmic calls with HF <3%. Red "X" represents the heteroplasmic variants called by MTool box, and blue "+"represents the heteroplasmic variants further filtered using deepSNV.

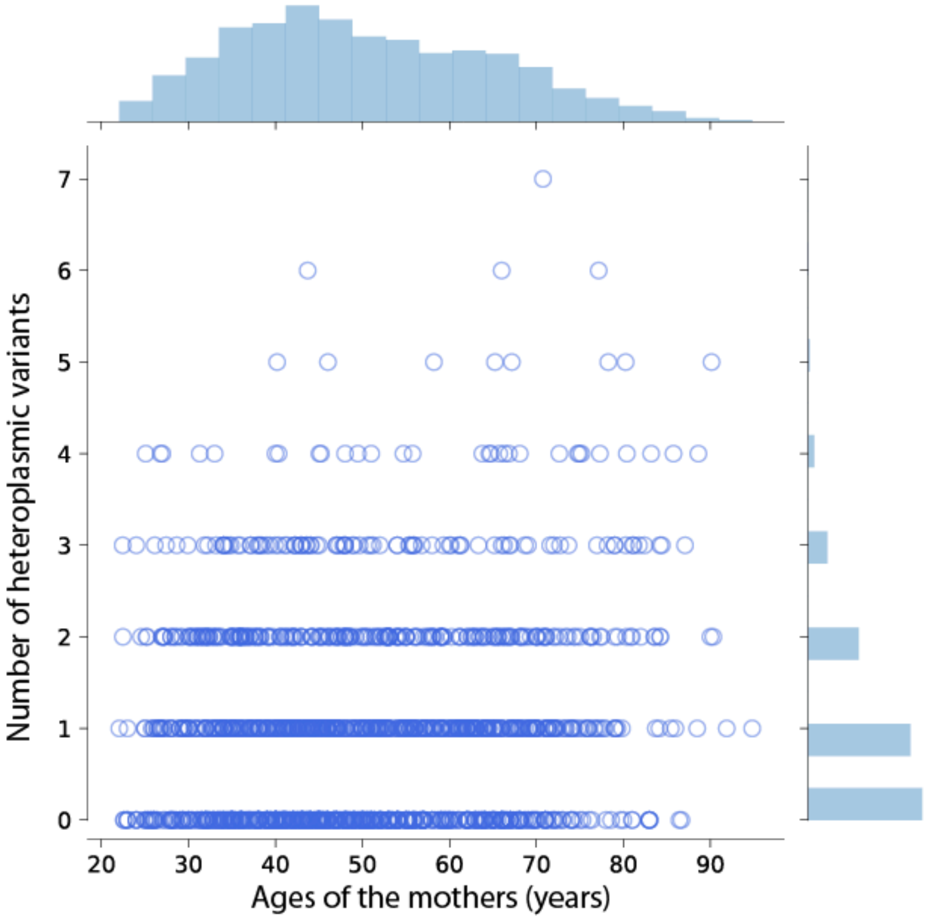
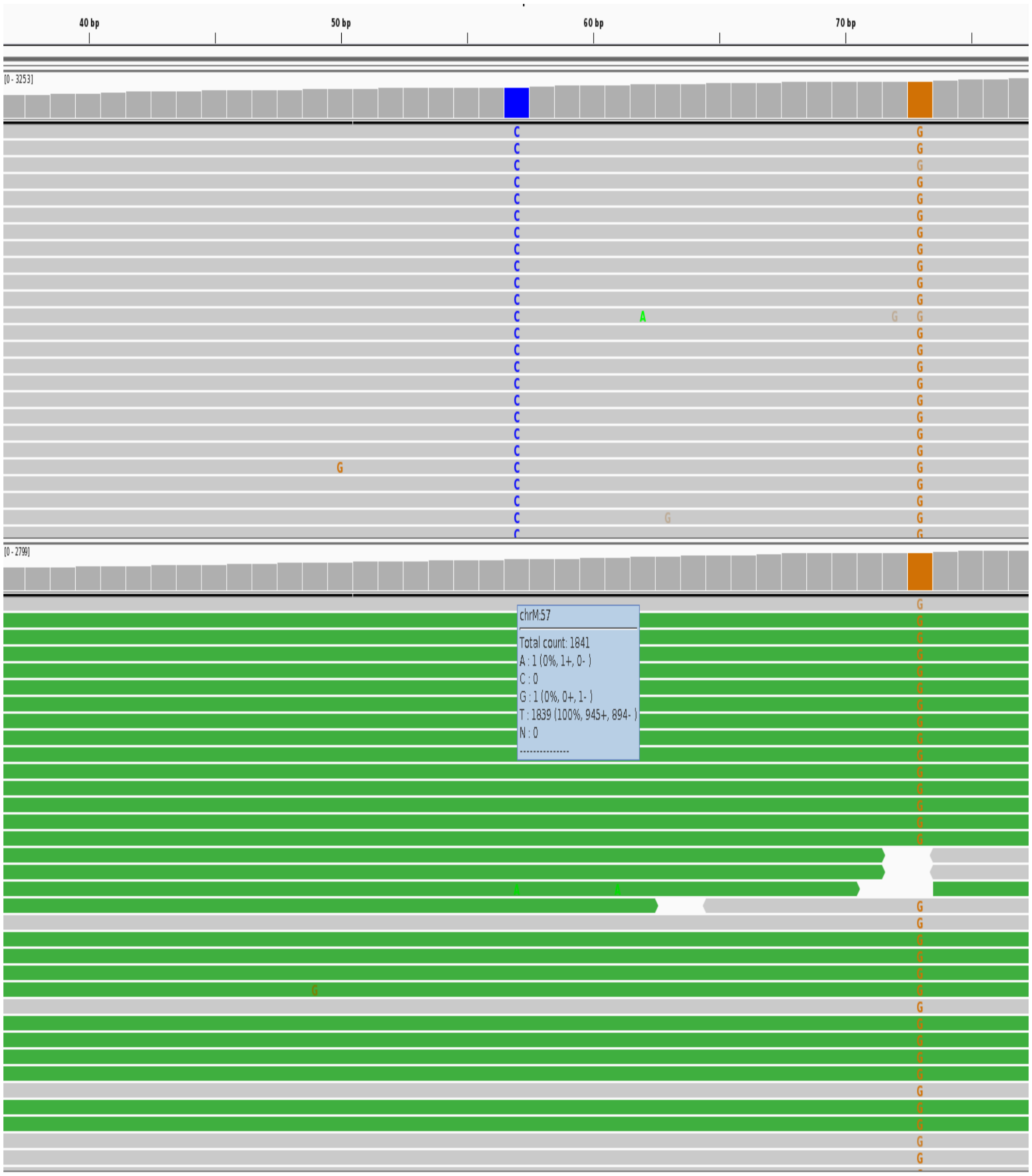
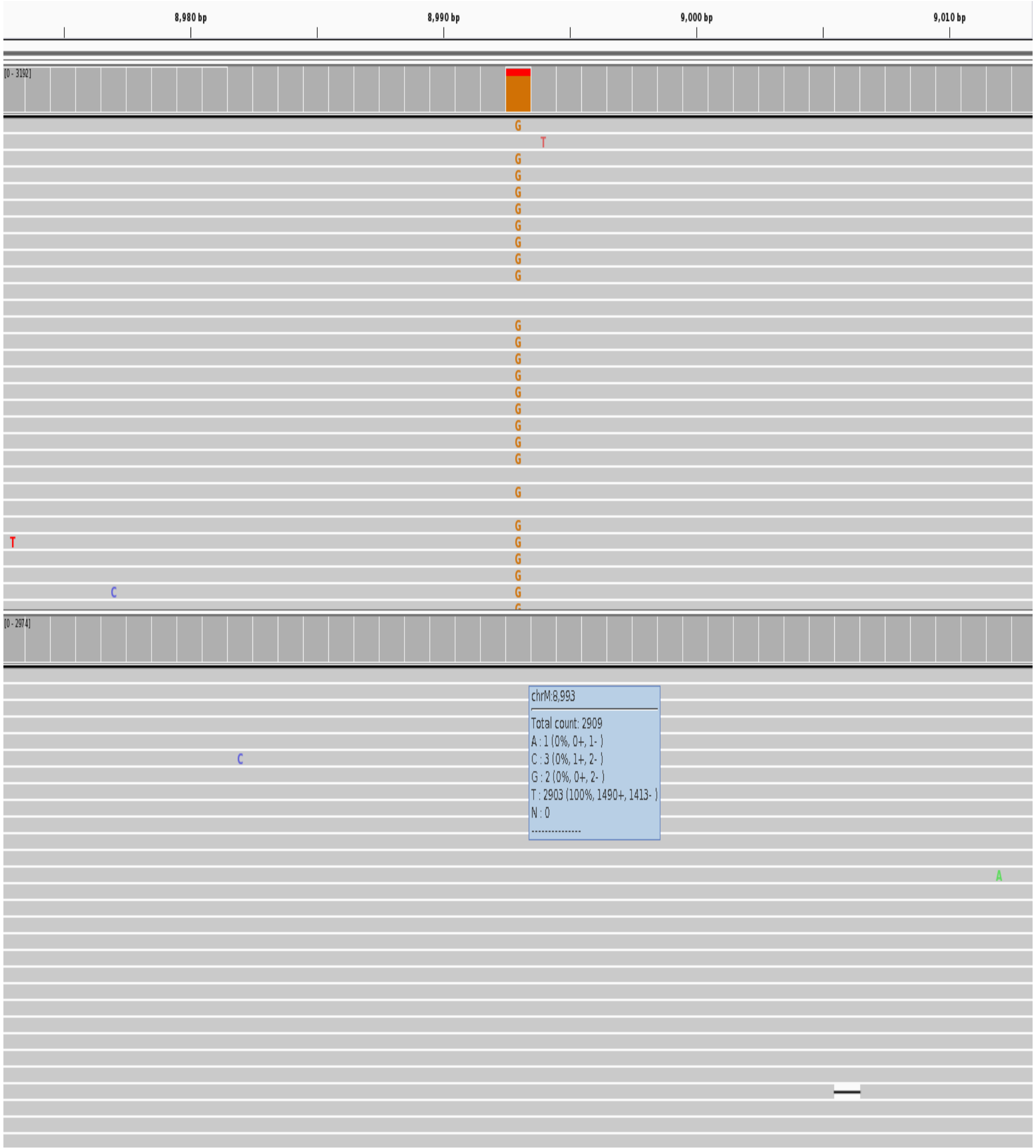


Fig. S3. Correlation of number of heteroplasmic variants in the mothers and the age of the mothers (R^2=0.17, P=6.42 x 10^{-11}, Pearson's correlation). The distributions of mother's ages and number of heteroplasmic variants are shown at the top and right side.



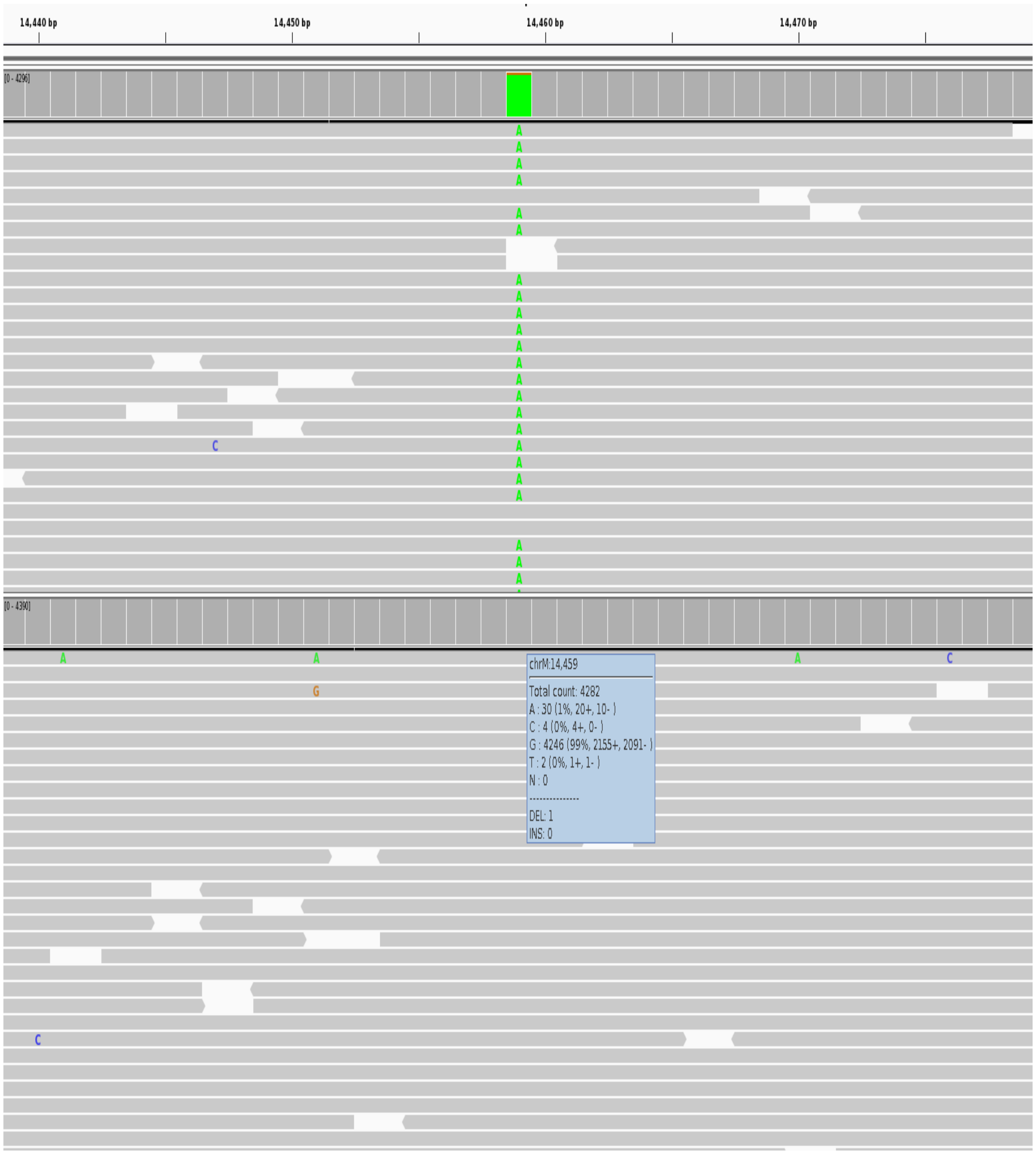
```
chrM:57
Total count: 1841
A : 1 (0%, 1+, 0- )
C : 0
G : 1 (0%, 0+, 1- )
T : 1839 (100%, 945+, 894- )
N : 0
---------------
```

Fig. S4. Read alignment views at mtDNA position 57. Top, offspring carrying 57 T>C variant; bottom, the corresponding mother carrying 57 T allele. The mother and offspring belonged to the same mtDNA sub-haplogroup, excluding a sample mix-up.


8,980 bp
8,990 bp
9,000 bp
9,010 bp
[0 - 3192]
G
T
C
[0 - 2974]
C
A

```
chrM:8,993
Total count: 2909
A : 1 (0%, 0+, 1- )
C : 3 (0%, 1+, 2- )
G : 2 (0%, 0+, 2- )
T : 2903 (100%, 1490+, 1413- )
N : 0
---------------
```

Fig. S5. Read alignment views at mtDNA position 8993. Top, offspring carrying 8993 T>G variant; bottom, the corresponding mother carrying 8993T allele. The mother and offspring belonged to the same mtDNA sub-haplogroup, excluding a sample mix-up.


14,440 bp
14,450 bp
14,460 bp
14,470 bp
[0 - 4296]
[0 - 4390]

```
chrM:14,459
Total count: 4282
A : 30 (1%, 20+, 10- )
C : 4 (0%, 4+, 0- )
G : 4246 (99%, 2155+, 2091- )
T : 2 (0%, 1+, 1- )
N : 0
---------------
DEL: 1
INS: 0
```

Fig. S6. Read alignment views at mtDNA position 14459. Top, offspring carrying 14459 G>A mutation; bottom, the corresponding mother carrying 14459G allele. The mother and offspring belonged to the same mtDNA sub-haplogroup, excluding a sample mix-up.

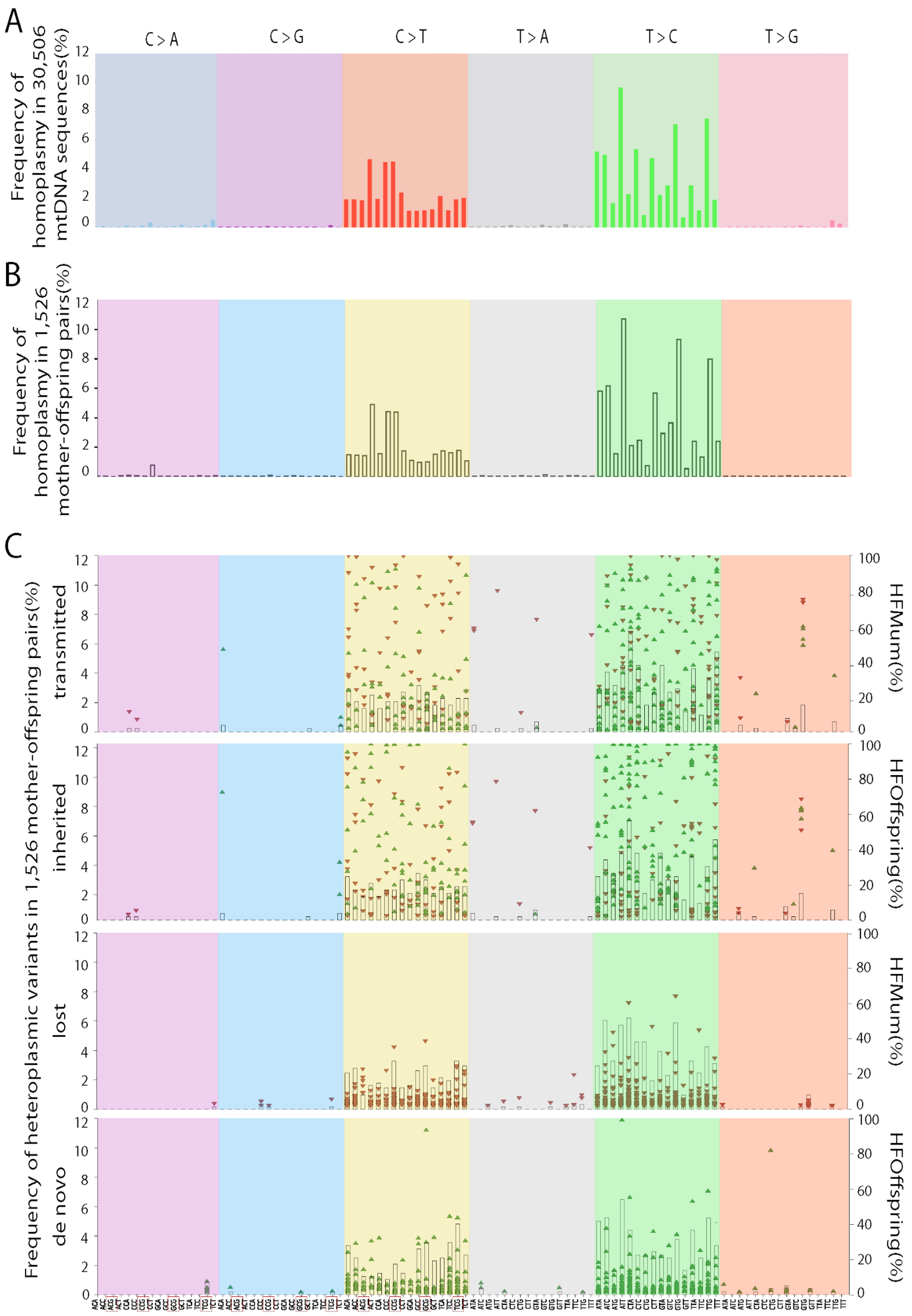


Fig. S7. Trinucleotide mutational signatures. (**A**) Trinucleotide mutational signatures of polymorphisms in 30,506 NCBI mtDNA sequences. (**B**) Trinucleotide mutational signatures of polymorphisms in 1,526 mother offspring pairs. (**C**) Trinucleotide mutational signatures of transmitted/inherited (the first two plots), lost and *de novo* heteroplasmic variants in 1,526

mother offspring pairs. HFs of the mothers are shown in the transmitted and lost trinucleotide mutational signatures. HFs of the offspring are shown in the inherited and *de novo* trinucleotide mutational signatures. The frequency bars for the six types of substitutions are displayed in different colours within different colour backgrounds labelled at the top of the graph. The trinucleotide signatures are on the horizontal axes at the bottom, and CpG islands are highlighted in the red rectangles. The frequencies of 96 possible mutation combinations are shown on the left side of y- axis, and HFs of the mother and offspring are shown on the right side of y-axis. The red triangles represent decrease heteroplasmic variants, and the green triangles represent increase heteroplasmic variants.

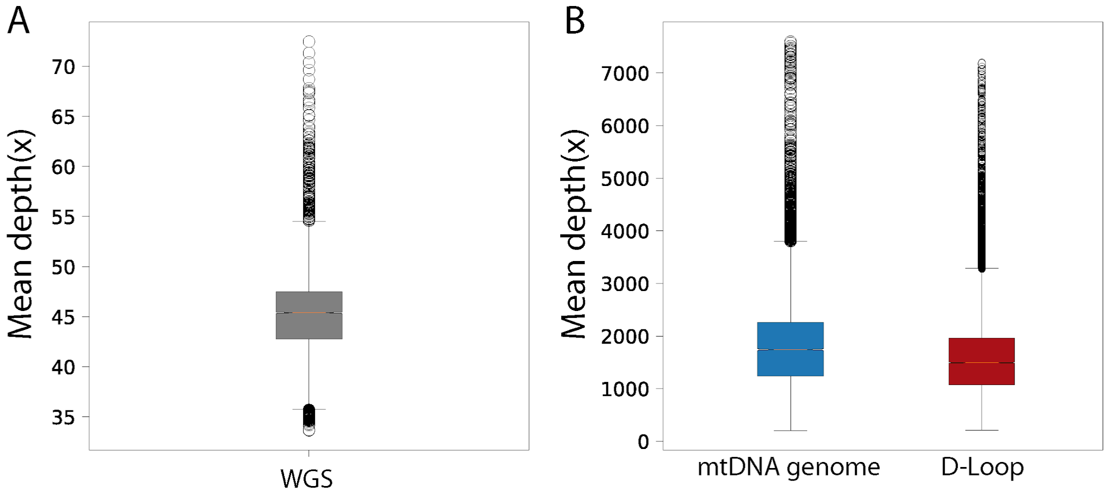


Fig. S8. (**A**) Box-whisker plots of mean depth of whole genome sequence from 12,975 individuals. (**B**) Mean depth of whole mitochondrial genome (left) and only D-loop region (right) from 12,975 individuals. The middle "box" represents the median, lower and upper quartile of the data. The upper and lower whiskers represent the data outside the middle 50%. The dots represent the data outside 25% - 75% of the data.

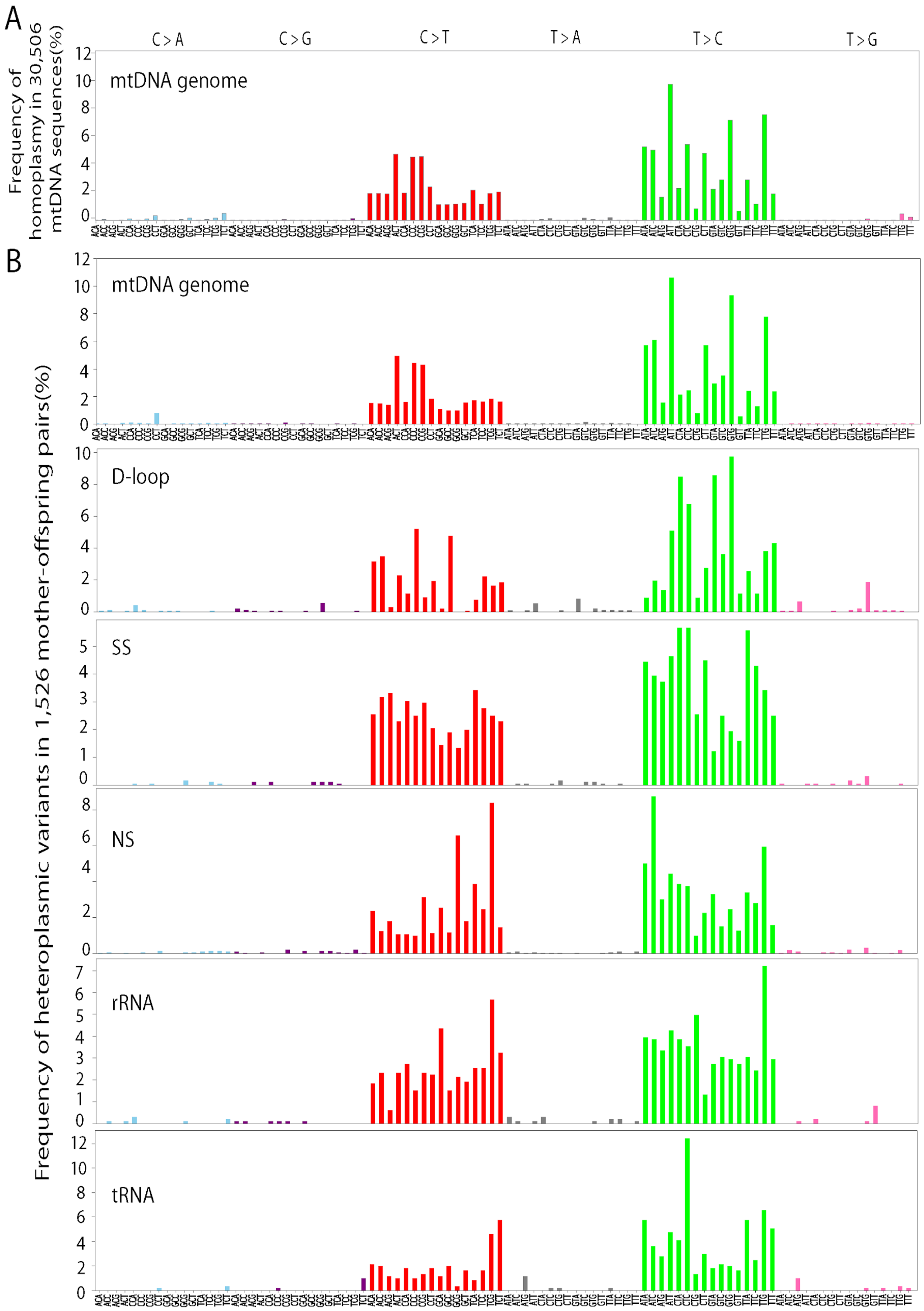
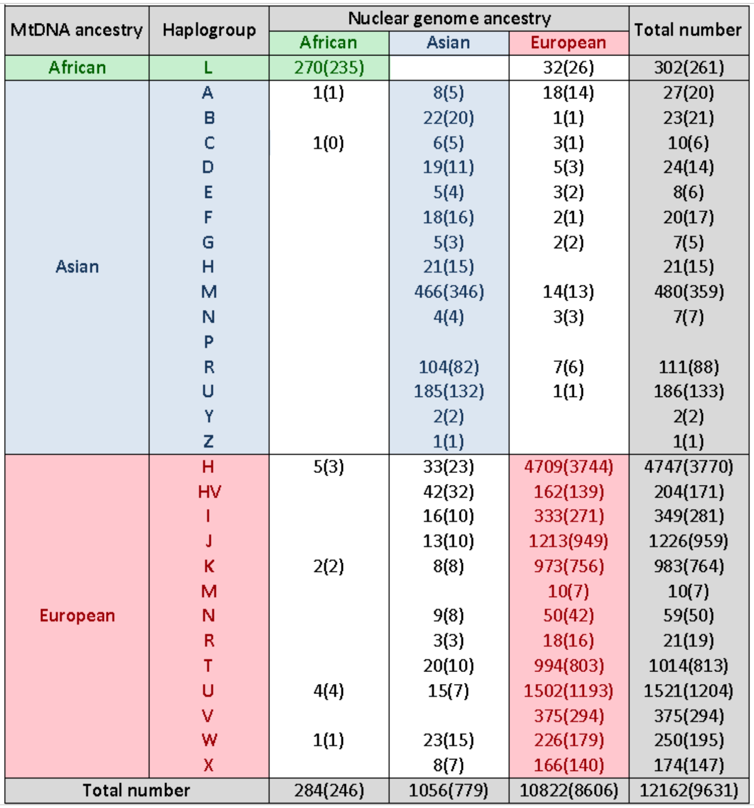


Fig. S9. Trinucleotide mutational signatures in 12,975 mtDNA sequences. (**A**) Trinucleotide mutational signatures of homoplasmic variants in 30,506 NCBI mtDNA sequences. (**B**) Trinucleotide mutational signatures of heteroplasmic variants in 12,975 mtDNA sequences. whole mtDNA genome, D-loop, SS, NS, rRNA and tRNA heteroplasmic variants are shown separately. The frequency bars for the six types of substitutions are displayed in different colours

labelled at the top of the graph. The trinucleotide signatures are on the horizontal axes at the bottom. The frequencies of 96 possible mutation combinations are shown on the left side of y-axis.



MtDNA ancestry	Haplogroup	Nuclear genome ancestry			Total number
		African	Asian	European	
African	L	270(235)		32(26)	302(261)
Asian	A	1(1)	8(5)	18(14)	27(20)
	B		22(20)	1(1)	23(21)
	C	1(0)	6(5)	3(1)	10(6)
	D		19(11)	5(3)	24(14)
	E		5(4)	3(2)	8(6)
	F		18(16)	2(1)	20(17)
	G		5(3)	2(2)	7(5)
	H		21(15)		21(15)
	M		466(346)	14(13)	480(359)
	N		4(4)	3(3)	7(7)
	P				
	R		104(82)	7(6)	111(88)
	U		185(132)	1(1)	186(133)
	Y		2(2)		2(2)
	Z		1(1)		1(1)
European	H	5(3)	33(23)	4709(3744)	4747(3770)
	HV		42(32)	162(139)	204(171)
	I		16(10)	333(271)	349(281)
	J		13(10)	1213(949)	1226(959)
	K	2(2)	8(8)	973(756)	983(764)
	M			10(7)	10(7)
	N		9(8)	50(42)	59(50)
	R		3(3)	18(16)	21(19)
	T		20(10)	994(803)	1014(813)
	U	4(4)	15(7)	1502(1193)	1521(1204)
	V			375(294)	375(294)
	W	1(1)	23(15)	226(179)	250(195)
	X		8(7)	166(140)	174(147)
Total number		284(246)	1056(779)	10822(8606)	12162(9631)

Fig. S10. Number of subjects in each ethnic group based on the nuclear and mtDNA ancestry. The ethnicity predicted using the nuclear genome labeled at the top, the mtDNA major haplogroups and the populations defined by mtDNA genome labeled on the left side. Africans in the matched group highlighted in green, Asians in blue and Europeans in red. The mismatched group is shown in the white background. The numbers of unrelated individuals are shown in the brackets.

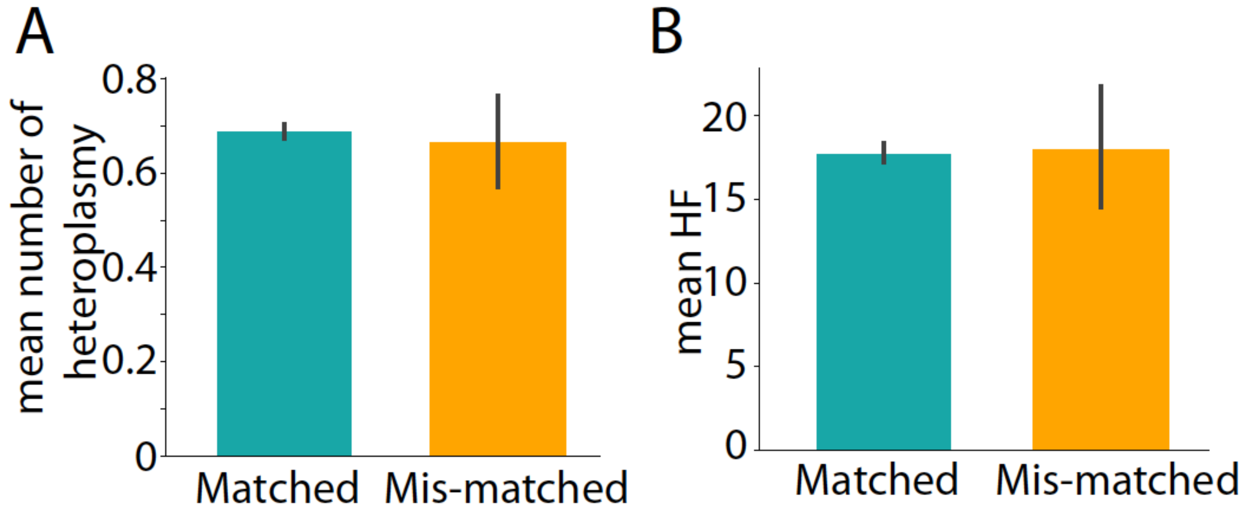
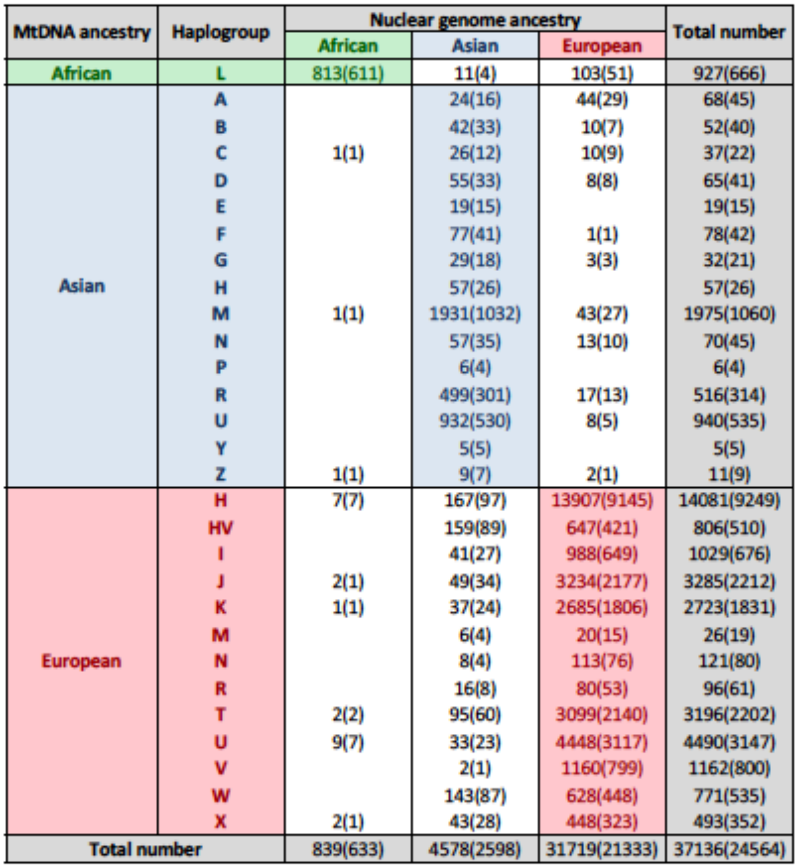


Fig. S11. (**A**) Mean number of heteroplasmic variants in the matched and mismatched groups. There was no significant difference between two groups (*P*>0.05, Wilcoxon rank sum test). (**B**) Mean HF in the matched and mismatched groups. There was no significant difference between two groups (*P*>0.05, Wilcoxon rank sum test). The error bar represents the standard deviation.



MtDNA ancestry	Haplogroup	Nuclear genome ancestry			Total number
		African	Asian	European	
African	L	813(611)	11(4)	103(51)	927(666)
Asian	A		24(16)	44(29)	68(45)
	B		42(33)	10(7)	52(40)
	C	1(1)	26(12)	10(9)	37(22)
	D		55(33)	8(8)	65(41)
	E		19(15)		19(15)
	F		77(41)	1(1)	78(42)
	G		29(18)	3(3)	32(21)
	H		57(26)		57(26)
	M	1(1)	1931(1032)	43(27)	1975(1060)
	N		57(35)	13(10)	70(45)
	P		6(4)		6(4)
	R		499(301)	17(13)	516(314)
	U		932(530)	8(5)	940(535)
	Y		5(5)		5(5)
	Z	1(1)	9(7)	2(1)	11(9)
European	H	7(7)	167(97)	13907(9145)	14081(9249)
	HV		159(89)	647(421)	806(510)
	I		41(27)	988(649)	1029(676)
	J	2(1)	49(34)	3234(2177)	3285(2212)
	K	1(1)	37(24)	2685(1806)	2723(1831)
	M		6(4)	20(15)	26(19)
	N		8(4)	113(76)	121(80)
	R		16(8)	80(53)	96(61)
	T	2(2)	95(60)	3099(2140)	3196(2202)
	U	9(7)	33(23)	4448(3117)	4490(3147)
	V		2(1)	1160(799)	1162(800)
	W		143(87)	628(448)	771(535)
	X	2(1)	43(28)	448(323)	493(352)
Total number		839(633)	4578(2598)	31719(21333)	37136(24564)

Fig. S12. Number of subjects in each ethnic group based on the nuclear and mtDNA ancestry in the validation dataset. The ethnicity predicted using the nuclear genome labeled at the top, the mtDNA major haplogroups and the populations defined by mtDNA genome labeled on the left side. Africans in the matched group highlighted in green, Asians in blue and Europeans in red. The mismatched group is shown in the white background. The numbers of unrelated individuals are shown in the brackets.

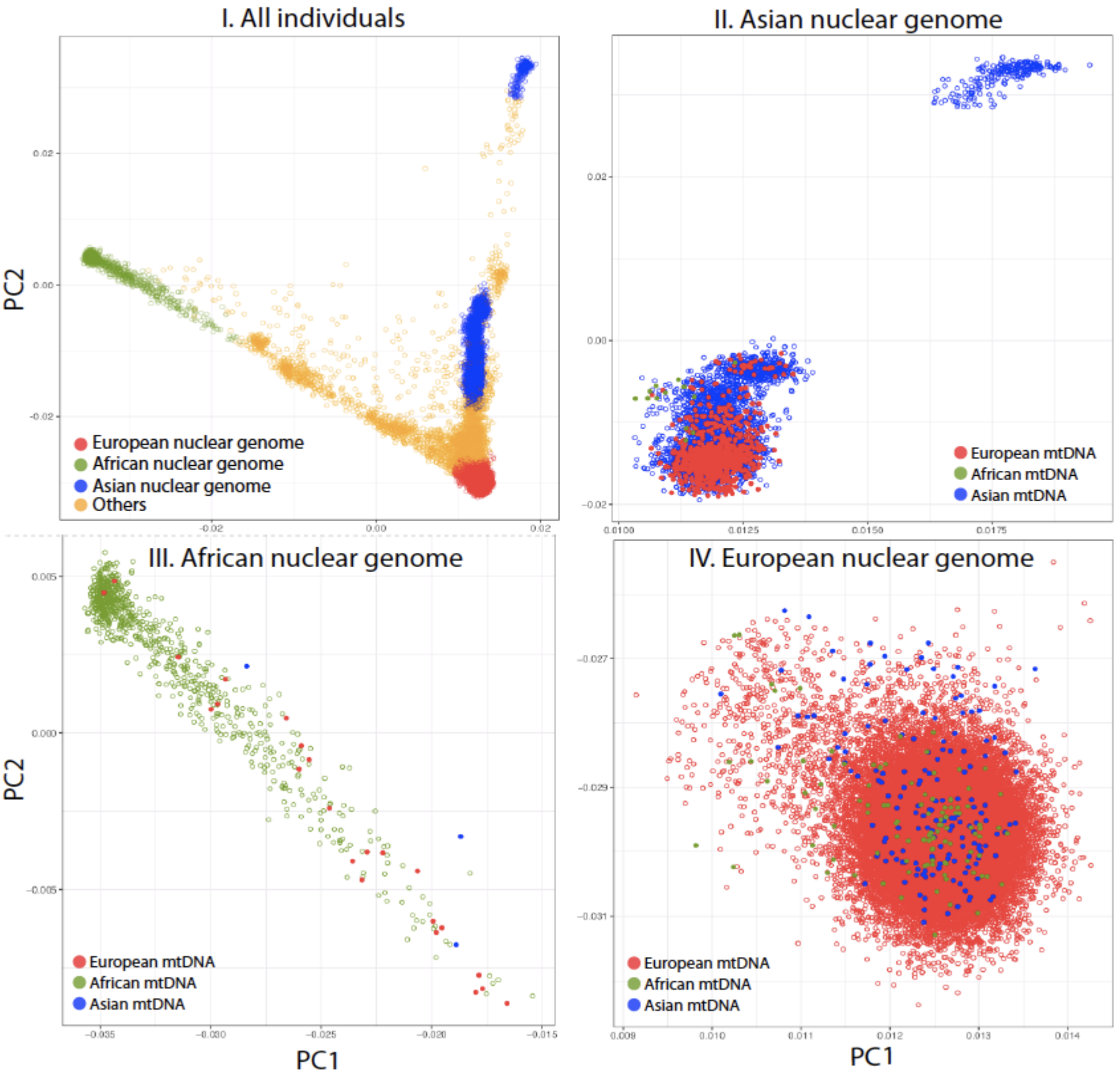


Fig. S13. Principal component analysis of the validation dataset. I. Projection of the nuclear genotypes at common SNPs onto the two leading principal components computed using the 1000
5 Genomes dataset, with individuals coloured by their assigned nuclear ancestry: Asian (blue), African (green), European (red) and Other (orange). The individuals coloured in blue, green and red in the boxes labelled II, III and IV are shown in panels II, III and IV, respectively, where they are coloured by their mitochondrial ancestries.

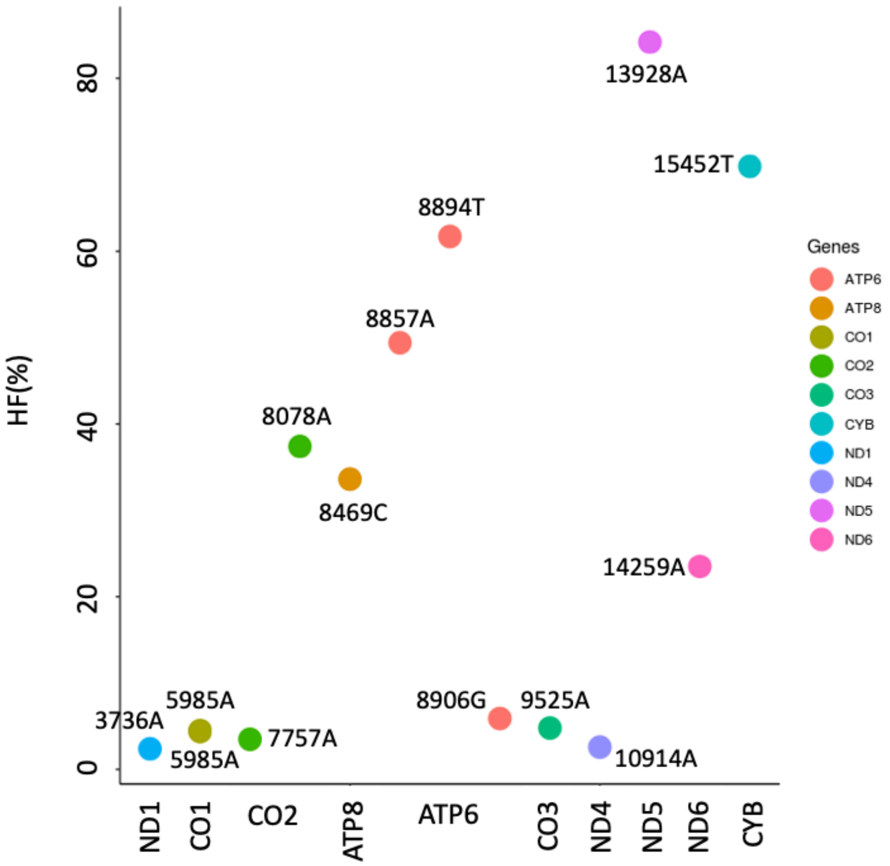


Fig. S14. Non-synonymous mtDNA variants found in people with a mis-match between their mtDNA and nuclear ancestry where the heteroplasmy matched the nuclear ancestry. HF = heteroplasmy fraction (%). X-axis = gene location of the variant.

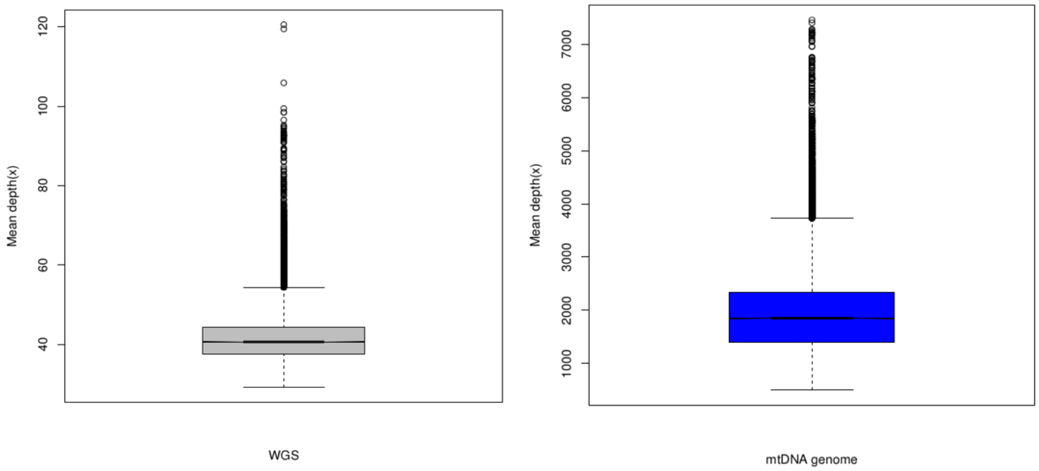


Fig. S15. Left: Box-whisker plots of mean depth of whole genome sequence from 40,325 individuals in the validation dataset. Right: Mean depth of whole mitochondrial genome from 40,325 individuals in the validation dataset. The middle “box” represents the median, lower and upper quartile of the data. The upper and lower whiskers represent the data outside the middle 50%. The dots represent the data outside 25% - 75% of the data.

Table S1. Summary of the heteroplasmic variants detected in the 1,526 mother-offspring pairs.

HF = heteroplasmic fraction (%).

		No. of individuals	No. of sample having heteroplasmic variant	No. of heteroplasmic variant	Average HF % (sd)
Mother		1,526	728 (47.8%)	1,043	11.4 (sd 12.2)
	Lost			614 (58.8%)	5.6 (sd 6.3)
	Transmitted			429 (41.2%)	19.5 (sd 13.9)
Offspring		1,526	648 (42.5%)	893	12.5 (sd13.0)
	de novo			477 (53.4%)	6.2 (sd 7.4)
	Inherited			416 (46.7%)	19.8 (sd 14.1)
Pair			929 (60.9%)		
Total		3,052		1,936	

Table S2. de novo rate in mtDNA genomic regions

Locus	Length (bp)	No. of *de novo*	*de novo* rate/base/genome	95% CI
DLOOP	1124	13	7.64E-05	(6.39E-05 - 9.06E-05)
MT-ATP6	681	19	1.83E-05	(1.10E-05 - 2.86E-05)
MT-ATP8	207	6	1.90E-05	(6.97E-06 - 4.13E-05)
MT-CO1	1542	35	1.49E-05	(1.04E-05 - 2.07E-05)
MT-CO2	684	21	2.01E-05	(1.25E-05 - 3.08E-05)
MT-CO3	784	17	1.42E-05	(8.28E-06 - 2.28E-05)
MT-CYB	1141	41	2.35E-05	(1.69E-05 - 3.19E-05)
MT-ND1	956	20	1.37E-05	(8.37E-06 - 2.12E-05)
MT-ND2	1042	19	1.19E-05	(7.19E-06 - 1.87E-05)
MT-ND3	346	11	2.08E-05	(1.04E-05 - 3.73E-05)
MT-ND4	1378	20	9.51E-06	(5.81E-06 - 1.47E-05)
MT-ND4L	297	5	1.10E-05	(3.58E-06 - 2.57E-05)
MT-ND5	1812	31	1.12E-05	(7.62E-06 - 1.59E-05)
MT-ND6	525	12	1.50E-05	(7.74E-06 - 2.62E-05)
MT-RNR1	954	20	1.37E-05	(8.39E-06 - 2.12E-05)
MT-RNR2	1559	28	1.18E-05	(7.82E-06 - 1.70E-05)
MT-tRNA	1508	39	1.69E-05	(1.21E-05 - 2.32E-05)

Table S3. Two-sample Kolmogorov-Smirnov test of heteroplasmic variants between different mtDNA regions

mtDNA regions	D	P value
Dloop vs non-syn	0.161	1.034E-06
Dloop vs rRNA	0.246	3.011E-08
Dloop vs tRNA	0.223	2.116E-05
syn vs non-syn	0.149	2.740E-05
syn vs rRNA	0.219	2.632E-06
syn vs tRNA	0.197	3.845E-04

Table S4. Numbers of observed heteroplasmic variants included in each nuclear and mtDNA ancestry group

Dataset	Group	Nuclear ancestry	Asian markers (N=1273)	European markers (N=941)
Observation dataset	**Matched**	Asian(646)	40	11
		European(8533)	408	309
	MisMatched	Asian(133)	14	5
		European(73)	6	5
Validation dataset	**Matched**	Asian(2110)	140	70
		European(21169)	1027	705
	MisMatched	Asian(491)	39	12
		European(185)	8	8

Note: Numbers of the samples in each ancestry group are shown in the brackets, column Nuclear ancestry; Numbers of the unique haplogroup specific variants on the mtDNA phylogenetic tree are shown in the brackets at the top.

Table S5. Disease groups and the sample size of each group included in this study

Disease Groups	Unaffected	Affected	Unknown	Total number	Note
BPD	41	1,123	0	1,164	Bleeding and platelet disorders
CNTRL	50	0	0	50	Process control samples (unaffected anonymous donors)
CSVD	10	145	91	246	Cerebral small vessel disease (stroke)
EDS	0	17	0	17	Ehler-Danlos (ED) and ED-like syndromes
GEL	2,199	2,619	0	4,818	100 000 Genomes Project - Rare Diseases Pilot
HCM	0	249	0	249	Hypertrophic cardiomyopathy
ICP	0	268	0	268	Intrahepatic cholestasis of pregnancy
IRD	0	715	0	715	Inherited retinal disorders
LHON	24	47	0	71	Leber hereditary optic neuropathy (mitochrondrial disease)
MPMT	37	532	0	582	Multiple primary malignant tumours
NDD	186	473	13	659	Neurological and developmental disorders
NPD	5	188	0	193	Neuropathic pain disorders
PAH	22	1,082	38	1,142	Pulmonary arterial hypertension
PID	350	984	14	1,348	Primary immune disorders
PMG	0	184	0	184	Primary membranoproliferative glomerulonephritis
SMD	117	125	13	255	Stem cell and myeloid disorders (Red and white blood cell disorders)
SRNS	15	236	0	251	Steroid resistant nephrotic syndrome
UKBio	0	763	0	763	Extreme red blood cell traits, RBC (Red Blood Cell count) and MCV (Mean cell volume) from UK Biobank
Total number	3,056	9,750	169	12,975	

Data S1. (separate file)

List of heteroplasmic variants detected in 1,526 mother-offspring pairs