An atlas of mitochondrial DNA genotype-phenotype associations in the UK Biobank

Ekaterina Yonova-Doing1,2,6, Claudia Calabrese3,4,6, Aurora Gomez-Duran3,4,5, Katherine Schon3,4, Wei Wei3,4, Savita Karthekeyan1, Patrick F. Chinnery3,4,7✉ and Joanna M. M. Howson1,2,7✉

Mitochondrial DNA (mtDNA) variation in common diseases has been underexplored, partly due to a lack of genotype calling and quality-control procedures. Developing an at-scale workflow for mtDNA variant analyses, we show correlations between nuclear and mitochondrial genomic structures within subpopulations of Great Britain and establish a UK Biobank reference atlas of mtDNA-phenotype associations. A total of 260 mtDNA-phenotype associations were new (\(P < 1 \times 10^{-5}\)), including rs2853822/m.8655 C>T (MT-ATP6) with type 2 diabetes, rs878966690/m.13117 A>G (MT-ND5) with multiple sclerosis, 6 mtDNA associations with adult height, 24 mtDNA associations with 2 liver biomarkers and 16 mtDNA associations with parameters of renal function. Rare-variant gene-based tests implicated complex I genes modulating mean corpuscular volume and mean corpuscular hemoglobin. Seven traits had both rare and common mtDNA associations, where rare variants tended to have larger effects than common variants. Our work illustrates the value of studying mtDNA variants in common complex diseases and lays foundations for future large-scale mtDNA association studies.

The 16,569-bp human mitochondrial genome has a compact genomic organization, with ~95% of the sequence encoding 13 proteins, 22 transfer RNAs and 2 ribosomal RNAs that are essential for oxidative phosphorylation (OXPHOS) and production of cellular energy in the form of ATP. Being maternally inherited, mtDNA undergoes negligible population level intermolecular recombination. As humans migrated out of Africa and populated the globe, they have acquired mitochondrial (mt) single-nucleotide variants (mtSNVs), which define geographical region-specific macro-haplogroups (related haplotypes)\(^4,5\). Many mtSNVs either directly affect mitochondrial function, or are in linkage disequilibrium (LD) with variants known to influence mitochondrial metabolism, and have been associated with common complex diseases including type 2 diabetes\(^7\) (T2D), cardiomyopathy and neurodegenerative disorders\(^6\).

Initial mtDNA association studies in complex traits were underpowered and yielded conflicting findings that were rarely replicated\(^6,11\). The inclusion of 265 mtDNA variants on the Affymetrix genotyping arrays used in the UK Biobank (UKBB)\(^12\), coupled with deep phenotypic data collected on half a million participants, provides an opportunity to address these issues and perform robust mtDNA genome-wide association studies (mtGWAS). Here, we establish a workflow for high-quality variant calling and imputation, and evaluate their role in 877 complex traits, providing a comprehensive atlas of mtSNV associations with diseases and endophenotypes in the UKBB population.

Results

Two hundred and sixty-five mtSNVs were genotyped in 488,377 UKBB participants (Supplementary Table 1). In the absence of standard procedures for quality control (QC) and imputation of mtSNVs from genotyping arrays, we adapted existing algorithms\(^13,14\) (Fig. 1 and Methods), to generate a quality-controlled set of 719 mtSNVs in 483,626 pan-ancestry individuals, referred to as the ‘full set’ (Supplementary Tables 2–4). Next, we defined a set of European (EUR) unrelated individuals (based on the kinship coefficient, removing first, second and third-degree relatives; \(N = 358,916\); Fig. 1, Supplementary Note and Methods). We excluded mtSNVs with minor allele count (MAC) < 10 or imputation INFO score < 0.7 (estimated by IMPUTE2; Fig. 1), resulting in 473 mtSNVs for association analyses in the unrelated EUR set (Supplementary Tables 5 and 6). In accordance with our power calculations (Supplementary Fig. 1), we restricted the analyses of binary traits to 416 mtSNVs with minor allele frequency (MAF) > 0.0001.

Calling and imputation of mtSNV genotypes. Genotype calling algorithms trained to detect three nuclear genotype clusters cannot reliably be used for mtSNVs where two clusters overwhelmingly predominate (Supplementary Fig. 2). To address this, we developed a four-stage genotype QC procedure (Methods and Supplementary Figs. 3–6): (1) pre-recalling QC, (2) manual recalling, (3) post-recalling QC and (4) imputation of mtSNVs not genotyped on the array (Fig. 1 and Supplementary Note). In stage 1, we excluded probe intensity outliers (Supplementary Figs. 3 and 4) and identified mtSNVs to recall in stage 2. We recalled 135 mtSNVs (50.9%) in stage 2 and, in stage 3, 248 high-quality mtSNVs passed genotype QC (Supplementary Tables 2 and 3 and Supplementary Fig. 5), increasing the mean per-sample call rate from 0.820 to 0.997 (Fig. 1 and Supplementary Table 2). In stage 4, we imputed 719 additional mtSNVs in the full UKBB set using 5,271 biallelic (homoplasmic) mtSNVs from a combined reference panel of 17,815 EUR, African and Asian mtDNA complete genomes (Fig. 1, Supplementary Fig. 5).
Fig. 1 | Mitochondrial genome PheWAS workflow. a, QC workflow depicts the steps taken to assure genotype quality: (1) pre-recalling QC, (2) manual recalling, (3) post-recalling QC and (4) imputation of mtSNVs not genotyped on the arrays. b, Examples of cluster plots of probe intensity for a mtSNV (m.14869 G>A) pre- and post-recalling genotyped in the ‘full set’ (N = 483,626 participants); the color legend corresponds to genotype assignment, with black dots indicating missing genotypes. c, Scatterplot showing the correlation of \(-\log_{10}\) MAFs of the 241 recalled mtSNVs compared to UKBB genotyped mtSNVs. The long-dashed lines indicate y = x and the short-dashed lines the linear regression fit. The gray shaded area represents the 95% confidence interval (CI) of the regression fit. Spearman’s correlation, two-sided P value (P = 1.8 × 10^{-205}) and rho correlation coefficient are provided. d, Scatterplots showing the correlation of \(-\log_{10}\) MAFs of the genotyped mtSNVs post-recalling (left) and the imputed variants (right) in UKBB mtSNVs compared to GenBank mtSNVs (MAC ≥ 30). Spearman’s correlation, two-sided P value (P = 8.6 × 10^{-45} for genotyped SNVs; P = 1.8 × 10^{-26} for imputed SNVs) and rho correlation coefficients are provided. UKBB individuals with nuclear-mitochondrial-matched African (AFR; N = 2,012 participants; green), Asian (AS, N = 888 participants; blue) and EUR (N = 358,916, unrelated participants; orange) ancestries were compared to corresponding GenBank genomes (EUR, N = 6,593; AFR, N = 704; AS, N = 3,587). e, CONSORT-like diagram showing the breakdown of participants and mtSNVs excluded at each step of the study. BT, binary trait; QT, quantitative trait.
mtDNA and nuclear genetic structure is correlated. In contrast to the structure of the nuclear genome in Great Britain (GB), less is known about the structure of mtDNA variation, and given the uniparental mtDNA inheritance, the two genomes are thought to be uncorrelated\(^{15}\). In the unrelated EUR dataset of the UKBB (358,916 participants), we observed differences in frequencies of \(\sim 2\%\) for macro-haplogroups J, W and I in three areas, Scotland, Northumberland/Tyne and Wear and Wales, compared to London (\(P = 7 \times 10^{-16}, 5 \times 10^{-9}, 1 \times 10^{-15}\) for J, W and I, respectively; Supplementary Tables 10 and 11). Among the sub-haplogroups with \(> 100\) carriers, that is, where we had statistical power to discern differences, we observed overrepresentation of one sub-haplogroup, J1b in Scotland (\(P = 3.5 \times 10^{-5}\); Extended Data Fig. 1).

We observed that 332 mtSNVs were associated (\(P < 5 \times 10^{-8}\)) with at least one of the first ten nuclear principal components (nucPCs; Extended Data Fig. 2). We also observed an association between macro-haplogroups I, J and K and nucPCs, specifically with clusters of individuals of Scottish, Northumbriam or Welsh ancestry (Fig. 2, Supplementary Fig. 7 and Supplementary Tables 12 and 13). We next tested whether the observed correlation of the macro-haplogroups with the nucPCs could be attributed to specific sub-haplogroups. The major sub-haplogroups (11a, 12a, 12f, 11b, 11c, K1a, K2a and K2b) were correlated with nucPCs, reflecting the findings with the macro-haplogroups, and H5a and H1b were additionally correlated. Sub-haplogroup distributions were comparable to those in the Wellcome Trust Case Control Consortium (WTCCC; Supplementary Table 9).

Our evaluation of both mtDNA principal components (mtPCs; Extended Data Figs. 2 and 3, Supplementary Fig. 8 and Supplementary Note) and nucPCs showed high correlation (average Spearman’s rho, 0.84) with the allele frequencies from the EUR ancestry components of three reference datasets (GenBank, 1000 Genomes and the Wellcome Trust Case Control Consortium cohort; Methods, Fig. 1 and Supplementary Table 7). Furthermore, the mtSNV imputation facilitated by the recalled variants together improved mtDNA haplogroup calling accuracy (median HaploGrep2 overall rank of 0.73, compared to 0.65 for pre-QC genotyped variants; Supplementary Fig. 6 and Supplementary Tables 8 and 9).

mtSNV–trait associations including both binary and quantitative traits and up to 473 mtSNVs. We found 88 mtSNVs were associated with one or more of 94 binary traits (Supplementary Table 17) at mitochondrial genome-wide significance (Bonferroni adjusted \(P < 5 \times 10^{-13} = 0.05/1000\) independent haplotypes; Methods), and 66 mtSNVs were associated with one or more of 27 quantitative traits (Tables 1–4, Figs. 3 and 4 and Supplementary Table 18). Two hundred and sixty mtSNV–trait associations were new, and all associations that reached the mtGWAS \(P\)-value threshold had a false discovery rate (FDR) < 5.6% (Supplementary Tables 17 and 18). We have used ‘lead’ variant throughout to describe the variant with the smallest association \(P\) value for a given trait.

We performed fine-mapping and found that for eight traits (mean platelet volume (MPV), plateletcrit (PCT), red blood cell count (RBC\#), mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), creatinine, estimated glomerular filtration rate (eGFR) creatinine (eGFR\(^{\text{cr}}\)) and height), associations with mtSNVs were due to multiple different mtSNVs using FINEMAP (Methods and Supplementary Table 19).

Finally, we tested for haplogroup-specific effects. Of the 22 binary traits tested, we found no evidence of haplogroup-specific effects for 18 traits, meaning that the lead variant (or variants in LD with the lead variant that were not genotyped or imputed in our study) captured the association thus making it unlikely that additional haplogroup-specific variants were responsible for the phenotype association. For 4 traits (ptosis of the eye, abdominal aortic aneurysm, volvulus and bladder problem), either the haplogroups or the lead variants modeled the association equally well (\(P > 0.001\); Methods and Supplementary Table 20). Of the 24 continuous traits tested, 4 were associated with haplogroups in addition to mtSNVs (cystatin C, cystatin C-derived eGFR, MCH and white blood cell count (WBC\#); Supplementary Table 20).

The observed associations were robust to sensitivity analyses assessing the analysis models used (Supplementary Information and Supplementary Tables 21 and 22). There was no significant evidence of association between mtDNA variants and age (\(P > 1 \times 10^{-8}\); minimum \(P = 0.003\) for association; minimum \(P = 0.002\) for interaction; Supplementary Tables 21–24). We found one variant \((rs2853516/m.3316G>A; MT-ND1: p.Ala4Thr)\) that was less common in men than women (MAF\(_{\text{men}}\) = 0.0018 versus MAF\(_{\text{women}}\) = 0.0025; \(P = 5 \times 10^{-5}\); Supplementary Tables 21–24). We observed an enrichment of associations with non-synonymous mtSNVs for binary traits (\(N = 14\); \(P = 6 \times 10^{-3}\)) but not for quantitative traits (\(N = 17\); \(P = 0.65\); \(P < 0.017 = 0.05/3\); Methods).

New mtSNV–multiple sclerosis associations. We identified three mtSNVs associated with multiple sclerosis (Table 1), including two new mtSNV associations tagging the K1a3 haplogroup (m.7559 A>G in MT-TD; OR (95% CI), 2.06 (1.59–2.67), \(P = 5.0 \times 10^{-4}\); and non-synonymous m.13117 A>G in MT-ND5; OR (95% CI), 1.65 (1.30–2.11), \(P = 4.2 \times 10^{-5}\)), and replicating a previously reported association with rs2853826/m.10398 A>G (\(P = 4.3 \times 10^{-5}\); Supplementary Note). We found that mtSNVs rather than the macro-haplogroups underpinned the association (\(P = 0.15\); Supplementary Table 20) resolving previous inconsistencies\(^{16–18}\).
African Americans\textsuperscript{20} and in white British\textsuperscript{7} individuals, but this variant was not available in the UKBB. We found no additional effect of the haplogroups on T2D risk in addition to rs2853822/m.8655 C>T ($P=0.58$; Supplementary Table 20).

mtSNVs associated with decreased height. Five mtSNVs tagging macro-haplogroup J, and a rare noncoding variant, rs267606617/m.1555 A>G (MAF = 0.002; \textit{MT-RNR1}), were associated with reduced height (Table 2 and Supplementary Table 18). Our fine-mapping analyses showed that rs28359172/m.12612 A>G (synonymous in \textit{MT-ND5}; MAF = 0.11) was independently associated with height (Supplementary Table 19) and this macro-haplogroup J-tagging variant was not correlated with the rare variant rs267606617 ($R^2=0.001$). The common variant (rs28359172/m.12612 A>G) was associated with a ~0.8-mm reduction in height, while the rare variant (rs267606617/m.1555 A>G), which has previously been associated with nonsyndromic deafness\textsuperscript{21–23}, was associated with ~4.3-mm reduction in height (Table 2). The effect sizes of these associations are comparable to the top 10\% of effects reported for the nuclear genome\textsuperscript{24}. Short stature is

![Fig. 2 | Distribution of the eight nuclear genome clusters and mtDNA haplogroups across Great Britain.](image)

The EUR unrelated individuals with birth coordinates ($N=327,665$ participants) were clustered based on the first ten nucPCs, resulting in eight nuclear clusters. The map of GB and Northern Ireland is colored according to the five regions identified by the most common clusters or combination of clusters in each region: (1) Scotland; (2) north of England (northeast and northwest); (3) north of England (Yorkshire and the Humber, and northwest of England); (4) south of England (Midlands, London, and southeast and south west of England); and (5) Wales. No data were available for Northern Ireland as participants were not recruited to the UKBB from Northern Ireland. The stacked bar charts represent the frequency of unrelated individuals in each of the eight identified nuclear clusters across eight EUR macro-haplogroups in each region (X, W, V, U, T, K, J and I). The macro-haplogroup H (the most common among EUR macro-haplogroups) was used as the baseline in the multinomial regression analysis and has been omitted. The white bars indicate the frequencies of individuals in the region and are used as a reference to compare the distribution of macro-haplogroups, that is, the area corresponding to the south of England. An asterisk denotes macro-haplogroups that are distributed differently (likelihood-ratio test, two-sided, $P<5 \times 10^{-5}$) between the south of England and the rest of the country. Color coding in ‘nuclear clusters’ refers to haplogroup frequency bar charts, while colors in ‘regions defined by main nuclear clusters’ mark the five regions of the country. The map was created using the GeoPandas package (https://geopandas.org/) with Python 2.7.
## Table 1 | New trait-mtSNV associations defined using UKBB ICD-10 and self-report codes

| Chapter | Trait | Definition | Locus | rsID; position (EA, EAF) | AA change | OR | 95% CI | P (FDR) | Haplogroup |
|---------|-------|-----------|-------|--------------------------|------------|----|--------|--------|------------|
| I. Infectious and parasitic diseases | Other specified bacterial agents | ICD-10: B968 | CO1 | rs879058417; 6,528 (T, 0.006)<sup>bc</sup> | p.Leu209Leu<sup>*</sup> | 1.36 | 1.18-1.57 | 3.1×10<sup>-5</sup> (0.04) | 65; K1a1b1f |
| II. Neoplasm | Skin of other and unspecified parts of face | ICD-10: C443 | ND3 | rs414879750; 10,084 (C, 0.007)<sup>bc</sup> | p.Ile9Thr | 1.36 | 1.17-1.57 | 4.7×10<sup>-5</sup> (0.05) | Homoplastic |
| | Intrathoracic lymph nodes | ICD-10: C771 | TQ | rs414563484; 4,336 (C, 0.002)<sup>bc</sup> | | 1.46 | 1.23-1.73 | 1.6×10<sup>-5</sup> (0.03) | H5a; U6d |
| | Descending colon | ICD-10: D124 | ND5 | rs1556642100;12,397 (G,0.002)<sup>bc</sup> | p.Thr21Ala | 1.92 | 1.41-2.64 | 4.4×10<sup>-5</sup> (0.05) | Homoplastic |
| III. Blood | Other iron deficiency anemias | ICD-10: D50.8 | TA | rs879226228; 5,633 (T, 0.008)<sup>bc</sup> | | 1.45 | 1.22-1.73 | 2.0×10<sup>-5</sup> (0.04) | J2b |
| | | | CYB | rs200336777; 15,812 (A,0.008)<sup>bc</sup> | p.Val356Met | 1.45 | 1.22-1.72 | 3.2×10<sup>-5</sup> (0.03) | |
| IV. Endocrine, nutritional and metabolic diseases | Hypokalemia | ICD-10: E87.6 | ND2 | rs367778601; 5,147 (A, 0.06)<sup>bc</sup> | p.Thr226Thr | 1.27 | 1.15-1.40 | 1.9×10<sup>-6</sup> (0.0005) | Homoplastic |
| | T2D | ICD-10; ICD9, 20002, 20003 | ATP6 | rs2853822; 8,655 (T,0.001)<sup>bc</sup> | p.Ile43Ile | 1.48 | 1.23-1.78 | 3.9×10<sup>-5</sup> (0.05) | ancestral variant common to L |
| VI. Nervous system | Multiple sclerosis | ICD-10: G35 | TD | rs1556423308; 7,559 (G, 0.005)<sup>bc</sup> | | 2.06 | 1.59-2.67 | 5.0×10<sup>-4</sup> (0.0003) | K1a3a |
| | | | ND3 | rs2853826; 10,398 (G, 0.21)<sup>bc</sup> | p.Thr114Ala | 1.15 | 1.07-1.23 | 4.3×10<sup>-5</sup> (0.05) | Homoplastic |
| | | | ND5 | rs878966690; 13,117 (G, 0.005)<sup>bc</sup> | p.Ile26Val | 1.65 | 1.30-2.11 | 4.2×10<sup>-5</sup> (0.05) | K1a3a |
| | Lesion of plantar nerve | ICD-10: G576 | CO3 | rs41482146; 9,667 (G, 0.01)<sup>bc</sup> | p.Asn154Ser | 1.49 | 1.24-1.80 | 3.2×10<sup>-5</sup> (0.04) | Usai Ib; Jlb2a |
| VII. Eye and adnexa | Ptosis of eyelid | ICD-10: H02.4 | CYB | rs193302994; 15,452 (A, 0.21)<sup>bc</sup> | p.Leu236Ile | 1.15 | 1.07-1.23 | 4.8×10<sup>-5</sup> (0.05) | J; T |
| IX. Circulatory system | Abdominal aortic aneurysm | ICD-10: I17.1 | DLOOP | rs369669319; 207 (A, 0.04) | | 1.38 | 1.18-1.61 | 4.6×10<sup>-5</sup> (0.04) | Homoplastic |
| X. Respiratory system | Spontaneous pneumothorax/ recurrent pneumothorax | 20002: 1126 | ATP8 | rs121434446; 8,393 (T, 0.008)<sup>bc</sup> | p.Pro10Ser | 1.67 | 1.33-2.11 | 1.4×10<sup>-5</sup> (0.02) | X2b |
| | | | TT | rs193303002; 15,927 (A,0.009)<sup>bc</sup> | | 1.70 | 1.35-2.15 | 8.0×10<sup>-4</sup> (0.02) | |
| XI. Digestive system | Bilateral inguinal hernia | K402 | ND1 | rs28357970; 3,796 (G, 0.015)<sup>bc</sup> | p.Thr164Ala | 1.43 | 1.22-1.67 | 1.2×10<sup>-5</sup> (0.02) | H1b1 |
| | | | Volvulus | K562 | ND1 | rs1599998; 4,216 (C, 0.21)<sup>bc</sup> | p.Tyr304His | 1.23 | 1.12-1.35 | 2.1×10<sup>-5</sup> (0.03) | J; T |
| | | | ND4 | rs86906886; 11,251 (G, 0.21)<sup>bc</sup> | p.Leu164Leu | 1.12 | 1.11-1.34 | 2.6×10<sup>-5</sup> (0.04) | |
| | | | CYB | rs193302994; 15,452 (A, 0.21)<sup>bc</sup> | p.Leu236Ile | 1.22 | 1.11-1.34 | 3.7×10<sup>-5</sup> (0.05) | |
| XII. Musculoskeletal system and connective tissue | Pain in joint (pelvic region and thigh) | ICD-10: M25.55 | CO1 | rs201617272; 5,913 (A, 0.01)<sup>bc</sup> | p.Asp4Asn | 1.47 | 1.25-1.74 | 4.0×10<sup>-4</sup> (0.01) | K1b |
| | Other shoulder lesions | ICD-10: M758 | RNR1 | rs200887992; 951 (A, 0.007)<sup>bc</sup> | | 1.74 | 1.38-2.19 | 2.9×10<sup>-6</sup> (0.008) | H2a1 |
| | | | Joint disorder | 20002: 1295 | ND2 | rs15566422875; 4,592 (C, 0.003)<sup>bc</sup> | p.Ile41Ile | 1.91 | 1.44-2.53 | 8.2×10<sup>-6</sup> (0.02) | H2a1a; Usaih |
| | | | ND4 | rs15564228398; 11,143 (T, 0.002)<sup>bc</sup> | p.Val144Ile | 2.07 | 1.53-2.81 | 2.5×10<sup>-6</sup> (0.007) | Cib811b |
| | | | CYB | rs147029798; 16,126 (C, 0.21)<sup>bc</sup> | | 1.12 | 1.06-1.17 | 1.3×10<sup>-5</sup> (0.02) | Homoplastic |

Continued
Table 1 | New trait–mtSNV associations defined using UKBB ICD-10 and self-report codes (continued)

| Chapter | Trait | Definition | Locus | rsID; position (EA, EAF) | AA change | OR | 95% CI | P (FDR) | Haplogroup |
|---------|-------|------------|-------|--------------------------|-----------|----|--------|---------|------------|
| XVIII. Symptoms, signs and abnormal findings | Polyuria | ICD-10: R35 | CYB | rs2835304; 14,793 (G, 0.05) | p.His16Arg | 0.83 | 0.76–0.91 | 4.6 × 10⁻⁵ (0.05) | USa; V2a1 |
| | Abnormal findings on diagnostic imaging of other parts of digestive tract | ICD-10: R93.3 | ND3 | rs41487950; 10,084 (C, 0.007) | p.Ile9Thr | 1.68 | 1.32–2.14 | 3.1 × 10⁻⁵ (0.04) | Homoplastic |

Summary of the single-variant mtDNA PheWas associations identified with P < 5 × 10⁻⁵ in the UKBB (up to 358,618 participants). Each variant had at least ten cases carrying the effect allele. Chapter: ICD-10 chapters; definition: ICD-10 or noncancerous illness self-reported diseases codes; locus: mtDNA-encoded gene; rsID: SNP ID as of dbSNP 153; position: mtDNA nucleotide position on rCRS (NC_012920); EA: effect allele; EAF: effect allele frequency; AA change: amino acid change; OR: odds ratio; 95% CI: 95% confidence interval; P value: the false discovery rate, FDR calculated with Benjamin–Hochberg procedure. Haplogroup, haplogroup(s) defined by the EA, according to Phylotree (Build 17); mtSNVs tagging more than two EUR haplogroups were reported as ‘homoplastic’. G, genotyped; I, imputed; M, mixed, that is, genotyped on one array only and imputed on the other; Ga, mtSNVs genotyped on one array only (that is, either the UKBB array or the UK BiLEVE Affymetrix Axiom array (UKBL) array) and not imputed on the other array (or excluded because of low INFO score on the other array).

Table 2 | mtSNV associations with height, airway function and longevity

| Trait | Locus | rsID; position (EA, EAF) | AA change | Beta | s.e. | P (FDR) | Haplogroup |
|-------|-------|--------------------------|-----------|------|------|---------|------------|
| Height | RNR1 | rs267606617; 1,555 (G, 0.002) | – | −0.060 | 0.020 | 4.3 × 10⁻⁵ (0.05) | NA |
| | ND5 | rs28359172; 12,612 (G, 0.011) | p.Val92Val | −0.010 | 0.003 | 2.0 × 10⁻⁴ (0.005) | J; K1a4c1 |
| Airway function (FeV1/FVC) | ND4 | rs3088053; 11,812 (G, 0.0001) | p.Leu351Leu | −0.001 | 0.0003 | 1.9 × 10⁻³ (0.03) | T2 |
| Longevity | Mother’s age | CO2 | rs3021089; 8,251 (A, 0.06) | p.Gly222Gly | 0.020 | 0.003 | 1.5 × 10⁻⁵ (0.03) | Homoplastic |
| | Parent’s age | CO2 | rs3021089; 8,251 (A, 0.06) | p.Gly222Gly | 0.020 | 0.003 | 1.6 × 10⁻⁵ (0.03) | Homoplastic |
| | CO2 | rs2853513; 16,223 (T, 0.07) | – | −0.020 | 0.003 | 9.6 × 10⁻⁶ (0.02) | Homoplastic |

Summary of the mtSNV associations with height (N = 358,045, participants), airway function (N = 266,818, participants) and longevity (up to 348,257 participants) traits, found at discovery P < 5 × 10⁻⁵. Parental age was inverse normal transformed within mothers and separately within fathers prior to analysis. EAF calculated on the set of samples with non-missing genotype/covariates; beta, effect size of the association; s.e., standard error; 95% CI, 95% confidence interval; P value represents the association of the EA with the listed trait; FDR, false discovery rate calculated with Benjamini–Hochberg procedure; ‘NA’ value indicates alleles that were not observed in Phylotree (Build 17); FEV1, forced expiratory volume in 1 s; FVC, forced volume capacity.

a well-recognized feature of inherited mitochondrial diseases. In some individuals, this has a neuroendocrine basis, but this is not the case in most people, in whom impaired cartilage-mediated growth has been implicated. Given that the known differences in OXPHOS and ATP synthesis are linked to different mtDNA haplogroups, the observed decrease in height might be linked to a less efficient ATP synthesis in individuals with haplogroup J, which can have an impact on poor growth and short stature, as previously described in patients with mitochondrial dysfunction. The lack of an effect of mtDNA haplogroups after including the two mtSNVs previously associated with longevity, however, we did not replicate published findings (P = 0.16; and rs28625645/m.489 T>C was absent in our dataset, R² < 0.2 with the available variants). Longevity has been associated with macro-haplogroup J previously; however, we found the association unconvincing in the UKBB (P = 0.05). Our analyses showed there were no haplogroup-specific effects on longevity (P = 0.13; Supplementary Table 20).

New mtSNV liver biomarker associations. We tested all eight liver function biomarkers available in the UKBB (Supplementary Table 16). We identified 23 mtSNVs associated with aspartate aminotransferase (AST) and 9 mtSNVs associated with alanine aminotransferase (ALT; P < 5 × 10⁻⁵; Supplementary Tables 18 and 19). The lead AST-associated variant, rs19330297/m.10238T>C (MAF, 0.03; MT-ND3), was associated with an increase of ~0.2 U l⁻¹ in AST and a missense variant rs193302980/m.14766T>C (MAF = 0.49) in MT-CYB was associated with an ~0.1 U l⁻¹ reduction in ALT (Table 3 and Supplementary Table 18). Notably, mtDNA variants have been associated with AST in the Japanese Biobank, but these were tagging the B4f haplogroup, which is absent in the EUR data set. Our haplogroup-specific analyses showed that either the haplogroups or the nonsynonymous variant rs193302980/m.14766T>C mtSNV could model the association with ALT, while for AST, there was no effect of the haplogroups in addition to the synonymous variant rs193302927/m.10238T>C (Supplementary Table 20), suggesting a variant not captured in this study may account for the AST association.

New mtSNV renal biomarker associations. Creatinine, cystatin C, urea and eGFR were associated with one or more of 16 mtSNVs
New mtSNV associations with blood cell traits. We identified 44 mtSNVs associated with at least one of 15 blood cell traits ($P < 5 \times 10^{-5}$; Fig. 4, Table 4 and Supplementary Tables 18 and 19). Many of the blood cell trait–mtSNV associations were shared across traits, for example, 29 were associated with both MCH and MCV (Supplementary Table 18). Our fine-mapping analyses identified multiple distinct mtSNV associations for five traits: RBC#, MCH, MCV, PCT and MPV (Table 4 and Supplementary Table 19). We found two mutations (rs199476112 and rs267606617) known to cause mitochondrial diseases that were associated with MCV and MCH (Table 4 and Supplementary Table 19). rs199476112/m.11778G>A (MAF, 0.004; MT-Nd4) is found in 95% of Europeans with Leber’s hereditary optic atrophy (LHON)38,39, while rs267606617/m.1555A>G (MAF, 0.002) in MT-RNR2 is associated with nonsyndromic deafness21-23 (in addition to our observed association with height). In both cases, the effect sizes were subtle ($\beta_{max}$, 0.3 s.d.). Interestingly, we found the MCV and MCH-associated variants, rs200336777/m.15812G>A (MAF, 0.008) and rs878944253/m.5633C>T (MAF, 0.008; Supplementary Table 18), that were also associated with anemia (Table 1). We identified an effect of haplogroup I in addition to the nonsynonymous mtSNV associations with MCH ($P = 4 \times 10^{-4}$; Supplementary Table 20).

Rare-variant and gene-based tests. The mtSNV–trait associations that we identified included 13 rare mtSNVs (MAF $\leq 0.01$). Across the seven traits (height, multiple sclerosis, AST, PCT, RBC#, MCV and MCH) that had both rare (MAF $\leq 0.01$) and common (MAF $> 0.01$) mtSNV associations, the rare variants tended to have larger effect sizes than common variants. For example, for multiple sclerosis, p.Thr114Ala in MT-Nd3 had an OR of 1.15, compared to the rare variants with an OR of 1.65–2.06 for p.Ile261Val in MT-Nd5 (MAF, 0.005) and mt.7559 in MT-TD (MAF, 0.005; Table 1). While for AST, common mtSNVs had $\beta$ from

### Table 3 | New mtSNV associations with serum biomarkers

| Trait | Locus | rsID: position (EA, EAF) | AA change | Beta | s.e. | $P$ (FDR) | Haplogroup |
|-------|-------|-------------------------|-----------|------|------|----------|------------|
| ALT   | CYB   | rs193302980; 14,766 (C, 0.49) | p.Thr7Ile | -0.008 | 0.002 | $2.8 \times 10^{-7}$ (0.001) | Homoplasic |
| AST   | ND3   | rs193302927; 10,238 (C, 0.03) | p.Ile60Ile | 0.03 | 0.005 | $1.3 \times 10^{-4}$ (2 $\times 10^{-4}$) | I |
| Creatinine | DLOOP | rs869183622; 73 (A, 0.45) | - | 0.007 | 0.001 | $8.9 \times 10^{-7}$ (0.003) | HV |
| Cystatin C | RNR2 | rs3928306; 3,010 (A, 0.26) | - | 0.009 | 0.002 | $2.9 \times 10^{-7}$ (0.001) | H1; J1 |
| eGFR$^{17}$ | DLOOP | rs869183622; 73 (A, 0.45) | - | -0.007 | 0.001 | $9.6 \times 10^{-7}$ (0.003) | HV |
| NDS   | DLOOP | rs2853499; 12,372 (A, 0.23) | p.Leu12Leu | 0.008 | 0.002 | $7.7 \times 10^{-7}$ (0.002) | U; K |
| eGFR$^{18}$ | RNR2 | rs3928306; 3,010 (A, 0.26) | - | -0.008 | 0.001 | $2.9 \times 10^{-7}$ (0.001) | H1; J1 |
| eGFR$^{19}$ | DLOOP | rs41402146; 462 (T; 0.09) | - | -0.01 | 0.002 | $1.2 \times 10^{-7}$ (0.0005) | H |
| Urea  | CYB   | rs2853504; 14,793 (G, 0.05) | p.His16Arg | -0.015 | 0.004 | $3.9 \times 10^{-8}$ (0.05) | U5a |

Summary of the 11 lead and distinct mtSNV–serum biomarker associations ($P \leq 5 \times 10^{-5}$) in the UKB (in up to 358,640 individuals). EA, effect allele; EAF, effect allele frequency; calculated with Benjamini–Hochberg procedure. eGFRcrcy, eGFR estimated using creatinine and cystatin C. P values for the lead mtSNV association for a given trait are highlighted in bold.
Table 4 | New mtSNV–blood cell trait associations

| Trait       | Locus          | rsID; position (EA, EAF) | AA change     | Beta   | s.e.  | P (FDR) | Haplogroup |
|-------------|----------------|--------------------------|---------------|--------|-------|---------|------------|
| MPV         | CO3            | rs2853493; 11,467 (G, 0.21) | p.Leu236Leu   | −0.0147 | 0.002 | 1.5 x 10^-9 (1.7 x 10^-9) | U; K       |
| PCT         | RNR1           | rs267606617; 15,555 (G, 0.002) | -             | −0.0829 | 0.002 | 7.1 x 10^-4 (0.001) | NA         |
|             | CO3            | rs2853825; 9,477 (A, 0.09) | p.Val91Ile    | −0.0206 | 0.003 | 1.8 x 10^-2 (2 x 10^-2) | U5         |
|             | CY8            | rs41518645; 15,257 (A, 0.02) | p.Asp171Asn   | −0.034  | 0.0061 | 3.1 x 10^-4 (1 x 10^-4) | J2; Klbl1a |
| RBC#        | CO3            | rs2853493; 11,467 (G, 0.21) | p.Leu236Leu   | −0.011  | 0.002 | 2.5 x 10^-7 (0.001) | U; K       |
|             | ND4            | rs199476112; 11,778 (A, 0.0004) | p.Arg340His   | −0.3306 | 0.05  | 4.0 x 10^-18 (6 x 10^-14) | T3; X2p1  |
|             | CY8            | rs41518645; 15,257 (A, 0.02) | p.Asp171Asn   | −0.0285 | 0.006 | 2.1 x 10^-4 (0.005) | J2; Klbl1a |
| MCV         | RNR1           | rs267606617; 15,555 (G, 0.002) | -             | 0.1427  | 0.02  | 3.4 x 10^-6 (8 x 10^-6) | NA         |
|             | ND4            | rs2853493; 11,467 (G, 0.21) | p.Leu236Leu   | 0.017   | 0.002 | 1.8 x 10^-5 (5 x 10^-5) | U; K       |
|             | ND5            | rs28359178; 13,708 (A, 0.13) | p.Ala458Thr   | 0.017   | 0.003 | 6.8 x 10^-11 (6 x 10^-11) | Homoplasmic |
| HCT         | ND4            | rs199476112; 11,778 (A, 0.0004) | p.Arg340His   | 0.0377  | 0.05  | 1.2 x 10^-19 (9 x 10^-15) | H; V       |
| MCH         | RNR1           | rs267606617; 15,555 (G, 0.002) | -             | 0.1399  | 0.02  | 1.2 x 10^-14 (2 x 10^-10) | NA         |
|             | ND4            | rs2853493; 11,467 (G, 0.21) | p.Leu236Leu   | 0.0185  | 0.002 | 5.4 x 10^-9 (5 x 10^-9) | U; K       |
|             | ND5            | rs28359178; 13,708 (A, 0.13) | p.Ala458Thr   | 0.0152  | 0.003 | 5.5 x 10^-9 (4 x 10^-9) | Homoplasmic |
| MHC         | RNR2           | rs2854128; 2,706 (A, 0.44) | -             | −0.0072 | 0.002 | 4.7 x 10^-8 (0.05) | H         |
| RDW         | TL2            | rs2853498; 12,308 (G, 0.23) | -             | −0.0138 | 0.002 | 3.5 x 10^-3 (3 x 10^-3) | U; K       |
|             | ND5            | rs2835499; 12,372 (A, 0.23) | p.Leu12Leu    | −0.0139 | 0.002 | 3.0 x 10^-11 (3 x 10^-11) | U; K       |
| EO#         | CY8            | rs41518645; 15,257 (A, 0.02) | p.Asp171Asn   | −0.0316 | 0.006 | 1.7 x 10^-7 (7 x 10^-4) | J2; Klbl1a |
| GRAN#       | CO1           | rs41413745; 6,734 (A, 0.01) | p.Met277Met   | 0.0349  | 0.008 | 4.3 x 10^-5 (0.05) | Homoplasmic |
| GRAN%(MYELOID) | DLOOP     | rs62581312;150 (T, 0.09) | -             | −0.0128 | 0.003 | 3.3 x 10^-4 (0.04) | Homoplasmic |
| LYMPH%#     | TL2            | rs2853498; 12,308 (G, 0.23) | -             | −0.01   | 0.002 | 1.5 x 10^-4 (0.004) | U; K       |
| WBC#        | CO3            | rs2853493; 11,467 (G, 0.21) | p.Leu236Leu   | −0.0103 | 0.002 | 1.4 x 10^-6 (0.004) | U; K       |
| MONO%O      | CY8            | rs41534044; 7,768 (G, 0.04) | p.Met61Met    | 0.0184  | 0.004 | 2.7 x 10^-4 (0.04) | U5b        |
| EO%         | CY8            | rs41518645; 15,257 (A, 0.02) | p.Asp171Asn   | −0.0258 | 0.006 | 2.0 x 10^-3 (0.03) | J2; Klbl1a |

Summary of the 28 lead and distinct mtSNV–blood cell trait associations (P < 5 x 10^-4) in the UKBB (in up to 345,714 individuals). EA, effect allele; EAF, effect allele frequency, calculated by using participants with non-missing genotype/covariates; AA, amino acid beta, effect size of the association; s.e., standard error; P value, association of the EA with the listed trait; FDR, false discovery rate calculated with the Benjamini–Hochberg procedure. Traits are as follows: MPV; PCT; RBC#; MCV; HCT, hematocrit; MCH; MHC; mean corpuscular hemoglobin concentration; RDW, red blood cell width; EO#, eosinophil count; GRAN#, granulocyte count; GRAN%MYELOID, percentage of granulocytes in the myeloid fraction; LYMPH%, lymphocyte count; WBC#; MONO%, percentage of monocytes; EO%, percentage of eosinophils. P values for the lead mtSNV association for a given trait are in bold.

Pathogenic mtSNVs occur on specific haplogroups. Our analysis also included six pathogenic mtSNVs causing mitochondrial diseases (nonsyndromic deafness: rs267606617/m.1555 A>G with MAF of 0.002; LHON: rs28616230/m.4171 A>G with MAF of 0.0006, rs199476112/m.11778 G>A with MAF of 0.0003, rs199476104/m.14484 T>C with MAF of 0.0008; mitochondrial encephalomyopathy, lactic acidosis and stroke-like episodes (MELAS) and Leigh syndrome: rs199476122/m.3697 G>A with MAF of 0.0002; myopathy: rs3879064421/m.14674 T>C with MAF of 0.00003)\(^1\) (Supplementary Note, Supplementary Fig. 10 and Supplementary Table 5), which typically have an incomplete clinical penetration and are associated with specific mtDNA haplogroups\(^1\). For example, patients with maternally inherited blindness due to mtDNA mutations causing LHON, overwhelmingly belong to macro-haplogroup H\(^4\), and patients with deafness due to the rs267606617/m.1555 A>G variant preferentially belong to macro-haplogroup H\(^4\). Although well established, these haplogroup associations are based on ascertaining individuals who have the disease. It is therefore not clear whether the haplogroup influences the clinical penetrance of the mtDNA mutation, whether the haplogroup predisposes to the mtDNA mutation or whether founder effects contribute to the associations.

We compared the carrier frequencies for the six mtSNVs in the EUR dataset, using macro-haplogroup H as reference (Supplementary Table 25). rs199476112/m.11778 G>A occurred equally across all nine EUR macro-haplogroups, but the odds of
carrying the rs199476104/m.14484 T>C on the J macro-haplogroup were lower (OR (95% CI), 0.31 (0.18–0.55), \( P = 6.4 \times 10^{-5} \)), while on the U macro-haplogroup, the odds were higher (OR (95% CI), 1.58 (1.38–1.80), \( P = 1.8 \times 10^{-11} \)). Similarly, the odds of carrying the rs267606617/m.1555 A>G on J and X were higher (OR (95% CI); 1.95 (1.79–2.13) and 3.29 (2.91–3.71), respectively; \( P < 1.0 \times 10^{-5} \) in both cases) and slightly lower on the T macro-haplogroup (OR (95% CI), 0.68 (0.58–0.78); \( P = 1.3 \times 10^{-4} \)). This adds to the evidence that haplogroup J influences the clinical penetrance of LHON, and excludes a single founder effect for each pathogenic mutation (Supplementary Fig. 9 and Supplementary Table 26). We did not analyze heteroplasmic mtDNA mutations because of unreliable allele calling.

Discussion
We have developed a workflow for QC, imputation and analysis of mtDNA genotypes that results in a set of variants of similarly high quality to the nuclear-encoded variants that will facilitate future mtDNA analyses. When applied to the UKBB, the workflow has provided a comprehensive reference dataset of mtDNA variant–trait associations to date, highlighting 260 new mtDNA–phenotype associations (\( P < 1 \times 10^{-5} \)). In addition, the gene-based tests identified new associations between red blood cell traits and MT-ND genes encoding respiratory chain complex I.

Mitochondrial dysfunction has been observed in several of the diseases that were associated with mtSNVs in our analyses, such as multiple sclerosis\(^{44}\), T2D\(^{45}\) and abdominal aortic aneurysm\(^{46}\). Similarly, some of the associations are with conditions that are complications of mitochondrial or mitochondrial-related diseases, for example, volvulus in a patient with MELAS\(^{47}\), eyelid ptosis in chronic progressive ophthalmoplegia\(^{48}\) and bladder problems in multiple sclerosis\(^{49}\). Interestingly, a recent study reported a link between mitochondrial dysfunction and pneumothorax consistent with the association we observe\(^{50}\). We replicated some
previously reported disease associations and discovered many more, despite limited statistical power owing to the UKBB being a general population cohort and the resulting low prevalence of many diseases.

Sixteen, predominantly homoplastic and macro-haplogroup tagging variants (for example, rs28359172/m.12612 A>G), showed pleiotropic effects across blood cell traits, liver biomarkers and renal biomarkers (Supplementary Table 18). Four of the mtSNVs associated with liver biomarkers showed an opposite direction to the effect on creatinine, and sixteen mtSNVs associated with RBC# showed opposite directions of effect for platelet parameters. The associations we observed between individual mtSNVs and specific traits could either be directly due to the variants themselves (as for nonsynonymous or RNA variants) or reflect ‘tagging’ of the real functional variants that reside within the same mtDNA haplogroup or sub-haplogroup, and the lack of whole-mtDNA genome sequencing data in the UKBB currently precludes in-depth fine-mapping. However, our findings are supported by related clinical observations. For example, height and eGFR are known to be affected in severe inherited mtDNA diseases\(^2\)\(^5\)\(^2\)\(^5\), which is consistent with the idea that ‘extreme phenotypes’ caused by pathogenic mutations and milder quantitative phenotypes seen in the general population constitute a spectrum of genetic effects arising from the mitochondrial genome\(^5\)\(^2\)\(^5\). Mitochondria also play key roles in the liver and the kidney. Liver mitochondria are a central site for integration of metabolic processes (including the urea cycle) and for exchange of metabolic intermediates, which is linked to OXPHOS\(^5\)\(^3\)\(^5\)\(^4\). For example, AST and ALT, the classical liver function biomarkers, are mitochondrial enzymes that play a central role in amino acid metabolism and in the replenishment of intermediates of the tricarboxylic acid cycle within mitochondria\(^5\)\(^3\)\(^5\)\(^\^\)\(^5\), and their synthesis is energy dependent. The kidneys have the second-highest mitochondrial content and oxygen consumption after the heart\(^9\). Renal mitochondria power the active transport needed for ion

Fig. 4 | PheWAS association results for blood cell and cardiometabolic traits. Circular concentric Manhattan plots showing a summary of associations (two-sided, \(P < 5 \times 10^{-5}\)) between mtSNVs and traits related to cardiometabolic health and endocrine traits, including red blood cell and platelet traits (quantitative traits, in up to \(N = 325,670\) participants), iron deficiency anemia (\(N = 279,179\) participants), and associated binary traits belonging to circulatory (\(N = 279,179\) participants) and endocrine (\(N = 322,038\) participants) systems, according to ICD-10 codes (Table 1). The black dashed line denotes the mitochondrial genome multiple-testing-adjusted significance threshold (two-sided, \(P = 5 \times 10^{-5}\)), while the blue dashed line denotes the nuclear genome significance threshold (two-sided, \(P = 5 \times 10^{-8}\)). mtSNVs passing the mitochondrial genome multiple-testing adjusted significance threshold were annotated by their locus, position and effect allele.
reabsorptions and play a role in nutrient sensing. Most liver and kidney pathologies are characterized by mitochondrial dysfunction and variants in a variety of nuclear-encoded mitochondrial genes have been linked to both mitochondrial and common complex diseases affecting the liver and kidneys.

Several mechanisms can explain the observed mtDNA associations. For some time, it has been known that common polymorphic mtDNA variants, in vertebrate and invertebrate model systems, regulate respiratory chain complex function, ATP levels, and calcium uptake. However, even subtle bioenergetic effects can have marked downstream consequences on cell mtDNA content, intramitochondrial transcription, and cell growth. Moreover, given emerging evidence that mitochondria act as metabolic hubs controlling diverse cellular processes, it is plausible that mtDNA variants also modulate canonical cellular signaling pathways in a tissue-specific manner, explaining why the same mtDNA variants are associated with different phenotypes, and sometimes with opposing effects. The experimental dissection of these mechanisms is a key next step in the interpretation of our findings.

Lastly, to our knowledge, this is the most comprehensive study exploring the population structure of mtDNA in GB to date. We showed that mtDNA structure is reflective of nuclear genetic ancestry and that macro-haplogroups J, K, I and W were more common in the Welsh, Northumbrians and the Scottish, probably reflecting known admixture from the Celts and Vikings (which is also apparent from analyses of the nuclear genome). The reasons for the associations are not clear, but given recent evidence that the nuclear genome appears to shape mtDNA evolution, it is tempting to speculate that maintaining nuclear–mitochondrial compatibility is driving this at a population level. However, we cannot exclude an ascertainment bias given the homogeneity of the EUR UKBB subset. We found no evidence that an interaction between mtDNA and the nuclear genome was associated with any of the quantitative traits. Our study also has some other limitations, including the known inaccuracy of self-reported phenotypes and a single time point for the cross-sectional data collection. However, the quantitative phenotype data provides new insights and forms the baseline for future prospective investigation.

In conclusion, understanding mitochondrial genetic architecture and the interaction between the nuclear and mitochondrial genomes will be important for reducing the burden of cardiometabolic and neurodegenerative diseases, among others. Our current findings establish the key role played by mtDNA variants in many quantitative human traits, and confirm their contribution to common disease risk. The atlas of UKBB mtSNV–trait associations provided here lays a firm foundation for future studies at the whole-mitochondrial genome level.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00868-1.

Received: 18 March 2020; Accepted: 7 April 2021; Published online: 17 May 2021

References
1. Saraste, M. Oxidative phosphorylation at the fin de siècle. Science 283, 1488–1493 (1999).
2. Giles, R. E., Blanc, H., Cann, H. M. & Wallace, D. C. Maternal inheritance of human mitochondrial DNA. Proc. Natl Acad. Sci. USA 77, 6715–6719 (1980).
3. Elson, J. L. et al. Analysis of European mtDNAs for recombination. Am. J. Hum. Genet. 68, 145–153 (2001).
4. Wallace, D. C. Mitochondrial DNA sequence variation in human evolution and disease. Proc. Natl Acad. Sci. USA 91, 8739–8746 (1994).
5. Wallace, D. C., Brown, M. D. & Lott, M. T. Mitochondrial DNA variation in human evolution and disease. Gene 238, 211–230 (1999).
6. Elson, J. L., Majamaa, K., Howell, N. & Chinnery, P. F. Associating mitochondrial DNA variation with complex traits. Am. J. Hum. Genet. 80, 378–392 (2007).
7. Poulton, J. et al. Type 2 diabetes is associated with a common mitochondrial variant: evidence from a population-based case–control study. Hum. Mol. Genet. 11, 1581–1583 (2002).
8. Wallace, D. C. & Chalkia, D. Mitochondrial DNA genetics and the heteroplasmacy conundrum in evolution and disease. Cold Spring Harb. Perspect. Biol. 5, a021220 (2013).
9. Keogh, M. J. & Chinnery, P. F. Mitochondrial DNA mutations in neurodegeneration. Biochim. Biophys. Acta 1847, 1401–1411 (2015).
10. Herrmann, C. & Howell, N. An evolutionary perspective on pathogenic mtDNA mutations: haplogroup associations of clinical disorders. Mitochondrion 4, 791–794 (2008).
11. Samuels, D. C., Carothers, A. D., Horton, R. & Chinnery, P. F. The power to detect disease associations with mitochondrial DNA haplogroups. Am. J. Hum. Genet. 78, 713–720 (2006).
12. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209 (2018).
13. Laurie, C. C. et al. Quality control and quality assurance in genotypic data for genome-wide association studies. Genet. Epidemiol. 34, 591–602 (2010).
14. Zhao, S. et al. Strategies for processing and quality control of Illumina genotyping arrays. Brief. Bioinform. 19, 765–775 (2017).
15. Yamamoto, K. et al. Genetic and phenotypic landscape of the mitochondrial genome in the Japanese population. Commun. Biol. 3, 104 (2020).
16. Hudson, G., Gomez-Duran, A., Wilson, I. J. & Chinnery, P. F. Recent mitochondrial DNA mutations increase the risk of developing common late-onset human diseases. PLoS Genet. 10, e1004369 (2014).
17. Kozin, M. S. et al. Variants of mitochondrial genome and risk of multiple sclerosis development in russians. Acta Natuara 10, 79–86 (2018).
18. Tranah, G. J. et al. Mitochondrial DNA sequence variation in multiple sclerosis. Neurology 85, 325–330 (2015).
19. Preste, R., Vitale, O., Clima, R., Gasparre, G. & Attimonelli, M. HmHvar: a new resource for human mitochondrial variations and pathogenicity data. Nucleic Acids Res. 47, D1202–D1210 (2019).
20. Mitchell, S. L. et al. Investigating the relationship between mitochondrial genetic variation and cardiovascular-related traits to develop a framework for mitochondrial phenome-wide association studies. BioData Min. 7, 6 (2014).
21. el-Schahawi, M. et al. Two large Spanish pedigrees with nonsyndromic sensorineural deafness and the mtDNA mutation at nt 1555 in the 12s rRNA gene: evidence of heteroplasmacy. Neurology 48, 453–456 (1997).
22. Casano, R. A. et al. Hearing loss due to the mitochondrial A1555G mutation in Italian families. Am. J. Med. Genet. 79, 388–391 (1998).
23. Bravo, O., Ballana, E. & Estivill, X. Cochlear alterations in deaf and growth retardation to mitochondrial diseases. J. Cell Biol. 2057–2066 (2019).
24. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. Hum. Mol. Genet. 27, 3641–3649 (2018).
25. Boal, R. L. et al. Height as a clinical biomarker of disease burden in adult mitochondrial disease. J. Clin. Endocrinol. Metab. 104, 2057–2066 (2019).
26. Holzer, T. et al. Respiratory chain inactivation links cartilage-mediated growth retardation to mitochondrial diseases. J. Cell Biol. 218, 1853–1870 (2019).
27. Gómez-Durán, A. et al. Oxidative phosphorylation differences between mitochondrial DNA haplogroups modify the risk of Leber's hereditary optic neuropathy. Biochim. Biophys. Acta 1822, 1216–1222 (2012).
28. Gómez-Durán, A. et al. Unmasking the causes of multifactorial disorders: OXPHOS differences between mitochondrial haplogroups. Hum. Mol. Genet. 33, 3534–3533 (2010).
29. Chen, A., Raule, N., Chomyn, A. & Attardi, G. Decreased reactive oxygen species production in cells with mitochondrial haplogroups associated with longevity. PLoS ONE 7, e46473 (2012).
30. Niemi, A.-K. et al. A combination of three common inherited mitochondrial DNA polymorphisms promotes longevity in Finnish and Japanese subjects. Eur. J. Hum. Genet. 13, 165–170 (2005).
31. Zhang, L. et al. Strikingly higher frequency in centenarians and twins of mtDNA mutation causing remodeling of replication origin in leukocytes. Proc. Natl Acad. Sci. USA 100, 1116–1121 (2003).
32. Niemi, A.-K. et al. Mitochondrial DNA polymorphisms associated with longevity in a Finnish population. Hum. Genet. 112, 29–33 (2003).
33. Santoro, A. et al. Mitochondrial DNA involvement in human longevity. *Biochim. Biophys. Acta* **1757**, 1388–1399 (2006).

34. Dato, S. et al. Association of the mitochondrial DNA haplogroup J with longevity is population specific. *Eur. J. Hum. Genet.* **12**, 1080–1082 (2004).

35. De Benedictis, G. et al. Mitochondrial DNA inherited variants are associated with successful aging and longevity in humans. *FASEB J.* **13**, 1532–1536 (1999).

36. Rose, G. et al. Paradoxes in longevity: sequence analysis of mtDNA haplogroup J in centenarians. *Eur. J. Hum. Genet.* **9**, 701–707 (2001).

37. Pacheu-Grau, D. et al. Mitochondrial antigens in personalized medicine. *Hum. Mol. Genet.* **22**, 1132–1139 (2013).

38. Jurkute, N. & Yu-Wai-Man, P. Leber hereditary optic neuropathy: bridging the translational gap. *Curr. Opin. Ophthalmol.* **28**, 403–409 (2017).

39. Wu-Yang-Ma, R. & Chinnery, P. F. Leber hereditary optic neuropathy. *J. Med. Genet.* **46**, 333–340 (2009).

40. Achilli, A. et al. Mitochondrial DNA backgrounds might modulate diabetes by altering mitochondrial function and intracellular mitochondrial signals. *PLoS Genet.* **5**, e1000474 (2009).

41. Achilli, A. et al. Mitochondrial DNA variants influence mitochondrial bioenergetics in *Drosophila melanogaster*. *Mitochondrion* **12**, 459–464 (2012).

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

43. Achilli, A. et al. The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. *Am. J. Hum. Genet.* **75**, 910–918 (2004).

44. Paternagni, S. et al. Mitochondria in multiple sclerosis: molecular mechanisms of pathogenesis. *Int. Rev. Cell Mol. Biol.* **318**, 49–103 (2017).

45. Achilli, A. et al. Mitochondrial DNA backgrounds might modulate diabetes complications rather than T2DM as a whole. *PLoS ONE* **6**, e21029 (2011).

46. Navas-Madroñal, M. et al. Enhanced endoplasmic reticulum and mitochondrial stress in abdominal aortic aneurysm. *Clin. Sci.* **133**, 1421–1438 (2019).

47. Hallac, A., Keshava, H. B., Morris-Stiff, G. & Ibrahim, S. Sigmoid volvulus in a patient with mitochondrial encephalomyopathy, lactic acidosis and stroke-like episodes (MELAS): a rare occurrence. *BMJ Case Rep.* **bcr2015213718** (2016).

48. Yu-Wai-Man, P. & Newman, N. J. Inherited eye-related disorders due to mitochondrial DNA variants. *Epigenetics* **7**, 17–27 (2012).

49. Fernández-Moreno, M. et al. Mitochondrial DNA haplogroups influence the risk of incident knee osteoarthritis in OAI and CHECK cohorts. A meta-analysis and functional study. *Ann. Rheum. Dis.* **76**, 1114–1122 (2017).

50. Kazuno, A. et al. Identification of mitochondrial DNA polymorphisms that alter mitochondrial matrix pH and intracellular calcium dynamics. *PLoS Genet.* **2**, e128 (2006).

51. Sissa, S. et al. Ancient mtDNA genetic variants modulate mtDNA transcription and replication. *PLoS Genet.* **5**, e1000479 (2009).

52. Salminen, T. S. et al. Mitochondrial genotype modulates mtDNA copy number and organismal phenotype in *Drosophila*. *Mitochondrion* **34**, 75–83 (2017).

53. Picard, M. et al. Progressive increase in mtDNA 3243A>T2DM as a whole. *PLoS ONE* **6**, e21029 (2011).

54. Compston, A. & Coles, A. Multiple sclerosis. *Lancet* **359**, 1221–1231 (2002).

55. Carstens, P.-O. et al. X-linked myotubular myopathy and recurrent spontaneous pneumothorax: a new phenotype? *Neurol. Genet.* **5**, e327 (2019).

56. Martín-Hernández, E. et al. Renal pathology in children with mitochondrial diseases. *Pediatr. Nephrol.* **20**, 1299–1305 (2005).

57. Stewart, J. B. & Chinnery, P. F. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nat. Rev. Genet.* **16**, 530–542 (2015).

58. Degli Esposti, D. et al. Mitochondrial roles and cytoprotection in chronic liver injury. *Biochem. Res. Int.* **2012**, 387626 (2012).

59. Houten, S. M. & Wanders, R. J. A. A general introduction to the biochemistry of mitochondrial fatty acid β-oxidation. *J. Inherit. Metab. Dis.* **33**, 469–477 (2010).

60. Owen, O. E., Kalhan, S. C. & Hanson, R. W. The key role of anaplerosis and cataplerosis for citric acid cycle function. *J. Biol. Chem.* **277**, 30409–30412 (2002).

61. Connors, T. M. et al. Mitochondrial DNA variants in mitochondrial diseases causing tubulointerstitial kidney disease. *PLoS Genet.* **13**, e1006620 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021
Methods

Study populations. UK Biobank. The UKBB study is a prospective population study (N = 502,682, age range of 40–69 years, predominantly Western Europeans) of UK residents who were recruited at 22 assessment centers across the United Kingdom between 2006 and 2010 (ref. 1). Participants attended a center, where they were deeply phenotyped, including various physical measurements, extensive health and lifestyle questionnaires and biological samples. DNA was extracted from buffy coat specimens at the UK Biocentre using a Promega Maxwell 16 Blood DNA Purification Kit (AS1010). Samples with sufficient DNA concentration and purity (as measured by a 260:280 ratio) were divided into aliquots, and 50µl of sample was shipped for genotyping at Affymetrix, where the majority of participants were genotyped using the Affymetrix Axiom array (UKBB). This is a customized genotyping array comprising 845,485 probesets for the assay of 820,967 SNVs and short insertions/deletions, including 243 mtSNVs.

A subset of the UKBB individuals (N = 50,008, male to female ratio of 1:1; white European ancestry; either never smokers or ‘heavy smokers’ (mean 35 packs per year) with lung function measurements) were selected to investigate the genetic determinants of smoking behavior, lung function and chronic obstructive pulmonary disease as part of the UKBB Lung Exome Variant Evaluation (UK BiLEVE) study4. As the UK BiLEVE participants are a subset of the UKBB study, DNA extraction, aliquoting and shipment procedures were the same as for the rest of the UKBB but UK BiLEVE participants were genotyped using the UK BiLEVE Affymetrix Axiom array (UKBL), which has >95% content overlap with the UKBB array, including 161 mtSNVs in common between the two arrays and 22 unique UKBB mtSNVs. A detailed description of the sample-based and variant-based QC procedures4 and genetic principal-component estimation performed using FLASHHPCA and the generated nuclear variants is available in the Supplementary Note. The UKBB received approval from the North West Multi-centre Research Ethics Committee.

Design of a bespoke workflow for quality control, recalling and imputation of mtDNA variants. A further, in-depth description of the workflow for QC, recalling and imputation of mtDNA variants is available in the Supplementary Note. The workflow uses the standard genotype call, intensity and array manifest files provided after genotyping, as well as publicly available whole-mitochondrial genomes for imputation purposes.

For stage 1, pre-recalling QC procedures were performed as follows: Within each genotyping array, we determined the per-mtSNV and per-sample call rates. We selected for recalling mtSNVs (Fig. 1) with call rates ≤0.990 (N = 7) and mtSNVs with lower call rates than in the 150,000 UKBB release (N = 8). Further, we excluded 54 samples that were outliers in the mean intensities over all SNVs, irrespective of batch, as well as 2,054 additional samples because of plate effects (Supplementary Figs. 3 and 4). Next, in the remaining individuals, variants with discordant MAF (>3%), compared to at least one of the reference datasets (GenBank, 1000 Genomes and WTCCC), were selected for recalling, only if they also showed suboptimal clustering. Finally, cluster plots for each variant were produced and visually inspected, and variants were selected for recalling on the basis of these plots (Supplementary Table 8 and Supplementary Fig. 1). This resulted in 135 poorly called mtSNVs from 265 mtSNVs genotyped in the UKBB.

For stages 2 and 3, the recalling procedures and post-recalling QC were performed as follows: Recalling of mtSNVs, either per array (N = 53, 39.3%) or per batch (N = 66, 48.9%), was completed using a bespoke validated R script (https://github.com/cody23/UKBB_mtSNV/tree/master/scripts/recalling). Variant cluster plots (per array and per batch) were generated and visually inspected and mtSNVs (N = 2; 1.5%) that showed overlapping clusters to an extent where recalling was not possible and monomorphic variants (N = 15; 10.7%) were excluded, resulting in a set of 248 mtSNVs (Supplementary Tables 1 and 2 and Supplementary Fig. 5). Finally, 2,643 samples with low call rates within each array (call rate ≤0.97) were excluded.

For stage 4, imputation (using IMPUTE2) was performed as follows: Before attempting imputation, we explored whether imputation of mtDNA variants was at all possible and tested haploid and diploid settings (Supplementary Note). We found no differences between the two settings and that, with 248 mtSNVs in common between the Biontogen and imputation sets, mtSNVs could be imputed with concordance >90% for both the major alleles and the minor alleles when the following cut-offs were implemented: INFO score ≥0.7 and MAC ≥10 (Supplementary Fig. 6).

We then imputed the UKBB dataset (Supplementary Note and Extended Data Fig. 2) against the 17,815 GenBank complete genomes and 5,271 biallelic SNVs. Imputation was performed separately on each array (N = 49,945 for UKBB and N = 438,377 for UKBL).

Haplogroup prediction. We predicted participants’ haplogroups using the command line version of the Haplogrep2 (v2.1.1) software8 on 483,626 UKBB individuals. In that part of the population, we included all 161 mtSNVs that were common between the genotyping array and Supplementary Table 8 and Supplementary Fig. 6). Predictions were successful for all 483,626 individuals (full set) when using the genotyped SNVs before QC (N = 265), for 462,366 individuals using only post-recalling SNVs (N = 248) and for 483,626 individuals (full set) using post-recalling and post-imputed SNVs (N = 768). For imputed SNVs, we selected those with an INFO score ≥0.7 in at least one of the two arrays and MAC ≥1 (N = 520), then imputed genotypes were set to hard calls with gooloo (https://www.well.ox.ac.uk/~creeman/software/gooloo/gooloo.html) conversion set to ‘ped’, with Haplotype Liability threshold ≥0.9. Haplogroup predictions were based on the rCRS-oriented PhyloTree (build 17)9 human phylogeny using VCF files as input. We also performed Haplogrep2 performed on 17,815 GenBank complete genomes, 2,419 1000 Genomes and 76 WTCCC individuals, using only the subset of homoplasmic variants identified by the MToolBox variant calling2 (Supplementary Table 9 and Supplementary Fig. 6).

To compare UKBB haplogroup predictions obtained with the three different reference mtSNV sets we used the Haplogrep2 overall rank score, a value ranging from 0.5 to 1 that reflects the ratio of observed and expected haplogroup-defining mtSNVs for that sample (Supplementary Note).

mtDNA variant annotation. We annotated the high-quality set of 473 minor alleles (including 248 genotyped and 225 imputed SNVs) identified in the UKBB EUR set using a programmatic API query of the HmtVar resource10. These included mitochondrial loci, variant type, AA change (for coding variants only), disease annotations in ClinVar and MITOMAP databases and several pathogenicity predictions and conservation scores (Supplementary Note, Supplementary Table 5 and Supplementary Fig. 11). We considered pathogenic variants with indication of association with diseases in both ClinVar and MITOMAP databases. We further selected variants with confirmed MITOMAP pathogenic status based on the list provided as of April 2018 (https://www.mitomap.org/OMAP/Confirmation/). We then retrieved the PhyloTree build 17 human phylogeny11 to retrieve the list of haplogroups tagged by minor alleles, and performed further integration using the MITOMAP list of top-level haplogroup markers (https://www.mitomap.org/OMAP/wiki/view/MITOMAP/HaplogroupMarkers/).

Modeling the geographical distribution of mitochondrial variation in Great Britain. To assess the distribution of mitochondrial variation in GB, as well as the relationship between the population structures of the two genomes, we first tested whether there was an association between mtSNVs and nucPCs (Extended Data Fig. 2). Then we calculated mtPCs using either all genotyped mtSNVs or genotyped+SNVs with MAC ≥10. mtPCs were computed with (R ≤0.2) or without pruning over the full set or the EUR set, and these were compared to similarly calculated mtPCs in the three reference datasets (Extended Data Fig. 3 and Supplementary Fig. 8). The correlation between the nucPCs and mtPCs in the UKBB was assessed (Extended Data Fig. 2).

Next, we added information on where participants were born (east-ness and north-ness of birth postcode) and attempted to predict those using mtSNVs (Supplementary Table 13). Based on the postcodes we assigned people to level 2 territory units of the Nomenclature of Territorial Units for Statistics (NUTS2) and assessed the distribution of macro-haplogroups across those units and whether any differences in frequencies were adjusted for by adding the first ten nucPCs as covariates. Finally, to understand the overlaps between the nuclear genome population structure and macro-haplogroups, we performed k-means clustering using the nucPCs and compared the distribution of these clusters and of the macro-haplogroups within the different NUTS2 territory units (Supplementary Tables 11 and 12 and Supplementary Fig. 7). Further, in-depth description of the analyses is available in the Supplementary Note.

Trait selection. To avoid underpowered analyses, we tested 767 categorical traits that had at least 500 cases each and MAF > 0.0001 (Supplementary Note, Supplementary Fig. 1 and Supplementary Table 14). Traits with low heritability, with less objective measurements or that were not fully available at the time we performed our analysis (for example, socioeconomic, lifestyle traits and magnetic resonance imaging phenotypes) were not analyzed. Analyzed traits included ICD-10 codes (N = 528), noncancerous illness self-reported diseases (category 2002; N = 166), health and medical history records (codes under category 100036; N = 12) and other remaining traits (N = 56), including combined categories (Supplementary Table 15). We tested 15 ICD-10 chapters, and diseases of the digestive system (chapter XI), musculoskeletal system and connective tissue (chapter XIII) and of the genitourinary system (chapter XIV) were the biggest portions of traits tested (~45% of all traits).

We also tested 126 quantitative traits, primarily from the metabolic and cardiovascular endophenotypes, including anthropometric traits, major lipids, blood pressure, blood cell traits and serum biomarkers, among others (Supplementary Note and Supplementary Table 16). We considered the thousands of continuous traits that were socioeconomic/lifestyle-related factors, for example, economic status, education, smoking, alcohol consumption and coffee drinking as these are more subjective measures and tend to be less heritable.

Determining a multiple-testing threshold for mtSNVs. As opposed to the nuclear genome, there is no conventional mitochondrial multiple-testing threshold. One could adopt a Bonferroni threshold over the total number of bases in the mitochondrial genome (16,569; P ≤ 10−5). Such a threshold is overly conservative as it does not recognize the known correlations between multiple
variants, which reflects the phylogenetic structure of mtDNA inheritance, nor the presence of known mutational cold spots. Previous candidate-variant studies corrected for number of variants tested, failing to account for correlation between variants, while studies focusing on haplogroups mostly used \( P < 0.05 \) (refs. 44, 114).

In a recent study, Kraja et al. performed a permutation analysis in one of the cohorts used in their meta-analysis and concluded that 49 SNVs represented the number of independent genetic effects. If we were to adopt similar strategies, we would arrive at \( P = 0.05 \times 473 = 1 \times 10^{-4} \) (EUR set) or \( P = 0.05 \times 20 = 2 \times 10^{-2} \) after LD pruning (\( R^2 = 0.2 \)). Both strategies do not take into account the actual number of independent variants one could test if whole-genome data were available and may be too liberal. Here we propose a Bonferroni threshold of \( P = 5 \times 10^{-3} < 0.05/1,000 \). This threshold reflects the number of distinct mtDNA haplotypes in the whole of the UKBB (\( N = 1,141 \); Supplementary Table 8) and is also equivalent to the number of independent variants (\( N = 1,036 \)) with MAC \( \geq 10 \) or \( R^2 < 0.2 \) within the GenBank reference panel. We calculated the Benjamin–Hochberg FDR across all the PheWAS associations to contextualize the reported \( P \) values for the association tests (\( N = 378,696 \); Supplementary Tables 17 and 18).

**Statistical analyses.** Unless stated otherwise, we performed all the statistical analyses using mtSNV allele dosages, calculated using QCTOOL version 2 (https://www.well.ox.ac.uk/~gav/qctool_v2/).

**Phenome-wide association studies.** For binary traits, we performed a two-sided Wald test for association between each trait and the mtSNVs using RVTEST\(^1\), adjusting for the effects of sex, array, and the first ten nucPCs. For quantitative traits, we performed a (two-tailed) \( z \)-score test and adjusted for trait-specific covariates (Supplementary Table 16). Additional analyses regarding the relationship between age, sex and mtSNVs are outlined in the Supplementary Note, and the effects of additional covariates used in the sensitivity analysis are presented in Supplementary Tables 21 and 22. We considered significant associations at two-sided \( P < 5 \times 10^{-5} \) (see above). Associations where none of the leading mtSNVs were in high LD (\( R^2 > 0.8 \)) with previously associated variants we classified as new. We calculated enrichment of specific types of mtSNVs by applying a two-tailed \( z \)-score test to compare the proportions of nonsynonymous, synonymous and noncoding variants within the leading associated mtSNVs over a background set (binary trait-associated mtSNVs, \( N = 416 \); quantitative trait-associated mtSNVs, \( N = 473 \)). Notably, the background distribution of mtSNVs did not change from that observed in the GenBank reference set of variants. All associations analyses were performed using mtSNV allele dosages.

**Statistical model assumptions.** Analyses of binary traits in the context of extreme case/control imbalance using logistic regression can be anti-conservative\(^{11} \). Therefore, we used both the score and Wald tests when testing mtSNVs for association with binary traits. For the 29 traits that were significant (\( P < 5 \times 10^{-5} \); MAC in cases \( \geq 10 \)), we further calculated the likelihood-ratio test to ensure robustness. We reported the Wald test results as they were the most conservative. The score and Wald tests were implemented in RVTEST, and STATA 14.2 (https://www stata.com stata14/) was used for the likelihood-ratio test. Finally, due to the modest number of variants tested, the correlation between variants and the high identity by descent (IBD) median ratio of shared mtSNVs (MAC > 0.01, \( R^2 < 0.2 \)) between pairs of individuals of 0.92 (minimum, 0.68) and 0.90 (minimum, 0.71) in GenBank and 1000 Genomes, respectively, we simulated inflation factor (lambda) distributions for the quantitative traits to test if they were in an acceptable range (Supplementary Note and Supplementary Fig. 11).

**For the interaction analysis,** we explored whether there was any evidence for age-by-mtSNV or sex-by-mtSNV interactions, by adding interaction terms to the regression models (STATA 14.2, likelihood-ratio test). For analysis of drugs affecting mitochondrial function, we tested whether the observed associations with quantitative traits were independent of the effects of drugs known to affect mitochondrial function such as antibiotics, drugs with heart or liver mitotoxicity or metformin\(^{11,12} \) (Supplementary Note).

To explore whether the mtSNV associations were independent of the effects of nuclear variants, we calculated, whenever possible, a nuclear polygenic score (nPGS) for each trait and modeled that together with the standard set of covariates described above. The nPGS was calculated for variants (\( P < 5 \times 10^{-5} \)) as follows:

\[
\text{ri} = \sum_{j=1}^{p} \beta_j x_{ij}
\]

Either conditionally independent or LD-pruned variants (\( R^2 < 0.2 \); PLINK 1.9 (ref. 13)) were used for nPGS calculations. The \( \beta \) estimates were derived from previous studies for blood cell traits\(^{12} \) and serum biomarkers\(^{13} \).

**Statistical fine-mapping.** We used Bayesian stochastic search approach using FINEMAP (v1.3)\(^{14} \) on the summary statistics from the single-variant analyses for fine-mapping of loci with multiple mtSNVs associations. Here we used default parameters – \( \sigma_{corr} \)-config and –\( \sigma_{corr} \)-group set to 0.7. Correlation between the mtSNVs required for the fine-mapping was calculated using LDSTORE\(^{15} \) and individuals from both UKBB arrays. We considered evidence for multiple signals present if log\( \text{odds} \), Bayes factor in favor of one or more underlying variants was \( >2 \).

We also performed sensitivity analyses using stepwise bidirectional regression on the individual-level data (cut-off for inclusion was set at the mtGWAS threshold: two-sided \( P < 5 \times 10^{-5} \) STATA version 14; https://www stata.com stata14/). Sensitivity analyses using conditional regression revealed broadly consistent results, with no trait having completely different mtSNVs selected by the two methods. Where inconsistencies occurred, these were usually because the conditional regression included an additional variant, or included variants correlated with (but not the same as) those from FINEMAP (Methods and Supplementary Table 19).

**Haplogroup analysis.** For all associated traits, we tested the distinct contributions of haplogroups and independently associated mtSNVs as defined by FINEMAP. We did this by comparing the model fit (by likelihood-ratio test) for the following three models:

1. Trait – covariates + mtSNV
2. Trait – covariates + macro-haplogroup
3. Trait – covariates + mtSNV + macro-haplogroup

Macro-haplogroups were modeled as a factorial variable with haplogroup H as the reference for all models.

**Haptype analysis.** We tested associations between known pathogenic mutations and haplotype backgrounds using multinomial logistic regression (‘multinom’ function of the nnet R package v.7.3-12; https://www.rdocumentation.org/packages/nnet/ versions/7.3-12/topics/multinom). We tested the nine EUR macro-haplogroups as the outcome and the pathogenic variants as predictors, using the H haplogroup as reference and adjusting for the effect of array, sex, the ten nucPCs and NUTS2-defined J and W geographical parameters (Supplementary Note).

**Gene-level analysis.** Variants with MAF \( \leq 0.02 \) were selected for gene-based testing. Score tests and covariance matrices for the mtDNA variants were generated for each phenotype using RVTEST\(^{14} \) with the meta-analysis command (meta). We performed gene-level analyses using the SKAT method (‘rareMAEts.range’ function) in the rareMAEts R package\(^{16} \) and presented two-sided gene-level \( P \) values. Mitochondrial coding and noncoding annotated features of rCRS, downloaded from HmtVar\(^{17} \), were selected for inclusion in the gene-based tests (Supplementary Table 5). We considered 46 features to group the mtSNVs as follows: 37 genes, 3 complexes (I, IV and V), all tRNAs together as a single feature, all tRNAs together as a single feature, the whole of DLOOP and its two hypervariable regions separately (HVSI and HVSI2) and all mtSNVs in noncoding regions as a single feature.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

Full summary statistics are provided at https://app.box.com/s/1w9eow90w9a3u60k40cep4km3jht/ and Zenodo (https://doi.org/10.5281/zenodo.4600973).

We used data from the following publicly available databases: https://www.mitomap.org/; https://www.ncbi.nlm.nih.gov/clinvar/; https://www.internationalgenome.org/.

**Code availability**

Code used to process UKBB data and source data used to generate the main figures are available at: https://github.com/clody23/UKBiobank_mtPheWas/.

Source data for Figs. 1–4 are available from: https://github.com/clody23/UKBiobank_mtPheWas/tree/master/source_files/Figure1/; https://github.com/clody23/UKBiobank_mtPheWas/tree/master/source_files/Figure2/; and https://github.com/clody23/UKBiobank_mtPheWas/tree/master/source_files/Figure3_and_4/.

Source data for Extended Data Figs. 1–3 are available from: https://github.com/clody23/UKBiobank_mtPheWas/tree/master/source_files/EDF1/; https://github.com/clody23/UKBiobank_mtPheWas/tree/master/source_files/EDF2/; and https://github.com/clody23/UKBiobank_mtPheWas/tree/master/source_files/EDF3/.

**References**

83. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12, e1001779 (2015).

84. Wain, L. V. et al. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. Lancet Respir. Med. 3, 769–781 (2015).

85. Surendran, P. et al. Discovery of rare variants associated with blood pressure regulation through meta-analysis of 1.3 million individuals. Nat. Genet. 52, 1314–1332 (2020).
Acknowledgements

We are grateful to G. Hudson and H. Griffin for discussions and the preliminary exploratory work that preceded this study; P. Surendran and T. Jiang for assistance with the genotype calling scripts; and W. Astle from the University of Cambridge for providing blood cell trait phenotypes and summary statistics. The BHF Cardiovascular Epidemiology Unit is supported by the UK Medical Research Council (MR/L003120/1), British Heart Foundation (RG/13/13/30194 and RG/18/13/33946) and the NIHR Cambridge Biomedical Research Center (BRC-1215-20014) and Health Data Research UK (which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome). P.C. is a Wellcome Trust Principal Research Fellow (212219/Z/18/Z) and a UK NIHR senior investigator, who receives support from the Medical Research Council Mitochondrial Biology Unit (MC_UU_00015/9), the Medical Research Council International Center for Genomic Medicine in Neuromuscular Disease, the Evelyn Trust and the NIHR Cambridge BRC (BRC-1215-20014) [*]. I.H. is funded by the British Heart Foundation (RG/13/13/30194) and the NIHR Cambridge BRC (BRC-1215-20014) [**]. E.Y.-D. was funded by the Isaac Newton Trust/Wellcome Trust/University of Cambridge Joint Research Grants Scheme. A.G.-D. is funded by the NIHR Cambridge BRC (146281). This research was conducted using the UKBB Resource under application numbers 20480, 7439 and 18794. [*]The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Author contributions

E.Y.-D. and C.C. performed analyses and drafted the manuscript. K.S. and A.G.-D. selected binary traits for analysis and filtering. W.W. performed GenBank data retrieval and initial QC. S.K. performed initial QC of UKBB data. P.F.C. and J.M.M.H. drafted the manuscript and supervised the work. All authors approved the final version of the paper.

Competing interests

J.M.M.H. and E.Y.D. became full-time employees of Novo Nordisk during the drafting of the manuscript. The remaining authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00868-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00868-1.

Correspondence and requests for materials should be addressed to P.E.C. or J.M.M.H.

Peer review information Nature Genetics thanks Valerio Carelli and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | Distribution of mitochondrial sub-haplogroups across Great Britain. The European unrelated individuals with birth coordinates (N = 327,665) were clustered based on the first 10 nucPCs, resulting in eight nuclear clusters. The map of Great Britain is colored according to the five regions identified by the most common clusters or combination of clusters in each region: (1) Scotland; (2) North of England (North East and West); (3) North of England (Yorkshire and the Humber, North West of England); (4) South of England (Midlands, London, South East and West of England); (5) Wales. No data were available for Northern Ireland. The stacked bar charts represent the frequency of unrelated individuals in each mitochondrial sub-haplogroups, in the five regions identified by the most common nuclear clusters or combination of nuclear clusters. The star indicates an over-representation (likelihood ratio test, two-sided P < 5 x 10^{-5}) of J1b sub-haplogroup in Scotland compared to the Midlands, London, South East and West region.
Extended Data Fig. 2 | Relationship between population structure in the nuclear and mitochondrial genomes. The figure shows (a) circular Manhattan plots of the association between the first 10 nucPCs and mtSNVs. For each mtSNV, the association was tested using a linear regression model: $Y = \beta_1 x X_1 + \beta_2 x X_2 + \beta_3 x X_3 + \beta_4 x X_4 + \beta_5 x X_5$ where $Y$ is a vector containing the values of a nucPC, $X_1$ is a vector of mtSNV dosages and $X_2-X_5$ are vectors containing covariate values (age, age squared, sex, and array) and $\beta_1-5$ represent the effect of each variable on the mean of $Y$. Wald test two-sided $P$-values are presented. The nucPCs are ordered from PC1 to PC10 from outside to in and black dots represent (Wald test, two-sided) $P < 5 \times 10^{-5}$. (b) 3D plots of the first three mtPCs; and (c) the relationship between the first three nuclear principal components (nucPCs, nucPC1 - left, nucPC2 - middle, nucPC3 - right) and the first two mitochondrial principal components (mtPCs). The latter were calculated using mtSNVs with MAF $> 0.01$ and $R^2 < 0.2$. The mtPCs in (a) and (b) were calculated using the following sets of genotyped mtSNV: (from left to right) all mtSNVs; mtSNVs with MAF $> 0.01$ only; and mtSNVs with MAF $> 0.01$ after LD-pruning at $R^2 < 0.2$. $N=$ the number of mtSNVs included in a given analysis. In (b) and (c) individuals are coloured according to macro-haplogroup carrier status.
Extended Data Fig. 3 | See next page for caption.
Extended Data Fig. 3 | Principal components analysis of the European set of UK Biobank participants in comparison to European participants in GenBank, 1000 genomes and WTCCC. Plots of the first three mitochondrial principal components (mtPCs) for individuals in: (a) the European set of UK Biobank (N = 358,916), (b) GenBank reference set used for imputation (N = 6,593), (c) 1000 Genomes individuals (N = 498) and (d) WTCCC controls (N = 747). For each of the three data sets, plots on the left-hand side show mtPCs calculated using pruned SNVs (R² < 0.2 for UK Biobank and R² < 0.1 for GenBank, 1000 Genomes and WTCCC) while the plots on the right were generated without LD-pruning. Individuals are colored according to macro-haplogroup carrier status. mtPCs were calculated using genotyped SNVs (MAF > 0.01).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Data were collected by UK Biobank and we downloaded data from the UK Biobank repository using UK Biobank helper programs [ukb_conv, ukbfetch, ukbgene, ukb_md5, ukb_unpack], July 2017 for genetic data, May 2018 for phenotype data, April 2019 for Biomarker data.

Data analysis

We used the following open source softwares (that are provided in the manuscript):
- FastQC v0.11.5 to check quality of fastq files generated with Illumina Hiseq.
- MToolBox v1.2 to perform mitochondrial variant calling on fastq files.
- Haplogrep2 v2.1.1 to perform haplogroup predictions.
- IMPUTE version 2 to perform imputation using 17,815 GenBank complete genotypes and 5,271 biallelic mitochondrial SNVs.
- Gotoo v0.7.5 to set genotypes in impute2 files to hard calls in ped files.
- GCTOOL v2 to perform mtsnv allele dosages and MAF calculation and all sorts of data parsing and file conversion on gen and bgen files.
- RVTESTS version 0.171009 to perform Wald and Score test for association between traits and mitochondrial SNVs, fitting different set of covariates depending on the trait. This tool was also used to generate meta-analysis files that were used as input for the gene level analysis.
- STATA v14.2 to perform likelihood ratio test for association between traits and mitochondrial SNVs, fitting different set of covariates depending on the trait. This tool was also used to perform fine-mapping of loci with multiple mitochondrial SNVs associations.
- PLINK v1.9 to calculate per mitochondrial SNV and per sample call rates, LD-pruning.
- FINEMAP v1.3 to perform fine-mapping of loci with multiple mitochondrial SNVs associations.
- LDSTORRE v1 to calculate correlation between mitochondrial SNVs used for fine-mapping.
- rareMETALS R package v6.8 to perform region-based meta-analysis using the SKAT method to compute a two-sided gene-level p-value.
- FlashPCA version 2 to define a genetically European group of UK Biobank participants using principal component analysis.
- reverse_geocoder Python module v1.5.1 to retrieve administrative regions (UK region, county and city), which were intersected with UK NUTS level 2 data, used as covariates for sensitivity analyses.
- Quantz v1.2.4 for power calculation
- https://github.com/cldry23/UKBiobank_mtmPheWas is a github repository that contains the R script used for variant recalling and several
other R and python scripts used for plotting
- R v3.4, R v3.5, python 2.7, python 3 to run custom R or python scripts

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data
All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

We have made the full summary statistics available through the Cambridge repository: https://app.box.com/s/vu9ewg9sv7ga3ua0skocq444xm3h; and on Zenodo: https://doi.org/10.5281/zenodo.4609973; We have also used data from the following publicly available data bases: www.mitomap.org; https://www.ncbi.nlm.nih.gov/clinvar/; https://www.internationalgenome.org/https://www.hmvar.uniba.it/. We provide the code used to generate main and supplementary plots, and variant calling in a Github repository: https://github.com/clodys23/UKBiobank_mtPheWas.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.
- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-list.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size
We used 488,377 UK Biobank participants and performed extensive variant and sample QC. We defined two groups of samples: 483,626 pan-ancestry individuals and a European set of 358,916 unrelated individuals. These are the maximum data set sizes available for any given analyses. Power calculations are provided in Supplementary Figure 1.

Data exclusions
We removed genotyping plates with problematic batch effects, intensity outliers, low sample call rate and non-Europeans and related individuals, exact numbers are provided in Figure 1. Individuals with outlying phenotype measurements were excluded as such individuals can potentially distort the statistical analyses and violate their assumptions. Details on sample exclusion are provided in Supplementary Table 15 and 16 for all the phenotypes tested. Exclusion criteria were pre-selected.

Replication
We did not attempt replication as this is the largest sample size available for these analyses to date.

Randomization
UK Biobank samples were randomised by sex and recruitment centre prior to genotyping.

Blinding
During genetic QC, analysts were blind to phenotype information. During phenotype QC, analysts were blind to genetic information. None of the authors took part in the data collection. We did not have access to group allocation information during data collection.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

| Materials & experimental systems | Methods |
|----------------------------------|---------|
| n/a                              | n/a     |
| ☒ Involved in the study          | ☒ Involved in the study |
| ☐ Antibodies                     | ☐ ChiP-seq |
| ☐ Eukaryotic cell lines          | ☐ Flow cytometry |
| ☐ Palaeontology                  | ☐ MRI-based neuroimaging |
| ☐ Animals and other organisms    |         |
| ☐ Human research participants    |         |
| ☒ Clinical data                  |         |
Human research participants

Policy information about studies involving human research participants

Population characteristics

The description of the UK Biobank, GeneBank, 1000 Genome and WTCCC cohorts used for this analysis is available in the "Study populations" Method's section of this manuscript and in the "GenBank", "1000 Genomes" and "Wellcome Trust Case Control Cohort" sections of Supplementary Information. UK Biobank participants used for association analyses were 358,916, all with mitochondrial and nuclear European ancestry (age range: 39-72 yrs, mean age = 57 yrs; 54% female). GenBank study participants were downloaded as mitochondrial complete genomes from NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/) (N=17,815), but no age and sex information are available. The nuclear ancestry was unknown for ~40% of the genomes, while mitochondrial ancestry was known for 100% of the genomes. Of the remaining GenBank genomes with known nuclear-mitochondrial ancestry, 37% were European, 33% Asian, and 30% African. Whole Exome Sequencing of 1000 Genomes individuals were downloaded from the 1000 Genomes project site (http://www.1000genomes.org/) (N=2,419), with information about sex (51% female). The 1000 Genomes participants come from 26 different populations from around the world (37% Asian, 28% African, 21% European, 14% Admixed American). No phenotype data has been collected for these participants. The WTCCC participants included in the study (N=763; 50% female; 98% with mitochondrial and nuclear European ancestry), living within England, Scotland or Wales ('Great Britain') and were sequenced with Illumina MiSeq. The cohort we used is part of the control individuals from the 1958 British Birth Cohort (58C) and individuals selected from blood donors recruited as part of the WTCCC project.

Recruitment

The description of how the UK Biobank, GeneBank, 1000 Genomes and WTCCC cohorts were recruited for this analysis is available in the "Study populations" Method's section of this manuscript and in the "GenBank", "1000 Genomes" and "Wellcome Trust Case Control Cohort" sections of Supplementary Information, as follows: The UKBB study is a prospective population study (N=502,682, age range: 39-72, predominantly western Europeans) of UK residents recruited at 22 assessment centers across the UK between 2006 and 2010. Participants attended a center, where they were deeply phenotyped, including various physical measurements, extensive health and lifestyle questionnaires, and biological samples. 17,815 complete GenBank human genomes were downloaded from the NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/) as explained in Wei W et al., 2017 (PMID: 29253894). Requirements for inclusion in the 1000 Genomes project were: individuals to be sequenced in the project had to have been explicitly consented for broad use and public distribution of extensive genotype or sequence data over the internet. Cell lines from the samples had to be available to the broad research community, to maximize the value of the 1000 Genomes data by facilitating follow-up studies and use of the sequence data to map cellular phenotypes. Further information about 1000 Genomes recruitment is available at The 1000 Genomes Project Consortium, 2010 (PMID: 20981002). The WTCCC cohort is part of a common set of 3,000 nationally-ascertained controls from England, Scotland and Wales. These controls come from two sources: 1,500 are representative samples from the 1958 British Birth Cohort and 1,500 are blood donors recruited by the three national UK Blood Services.

Ethics oversight

UK Biobank has approval from the North West Multi-centre Research Ethics Committee (MREC). The Wellcome Trust Case-Control Consortium subjects gave written informed consent and the project protocols were approved by the relevant research ethics committees in the UK. A list of investigators who contributed to the generation of the data is available from www.wtccc.org.uk. The 1000 Genomes consortium subjects provided their consent as explained in https://www.internationalgenome.org/sample_collection_principles.

Note that full information on the approval of the study protocol must also be provided in the manuscript.