

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31

Nuclear genetic control of mtDNA copy number and heteroplasmy in humans

Rahul Gupta^{1,2,3}, Masahiro Kanai^{2,3}, Timothy J. Durham^{1,2}, Kristin Tsuo^{2,3}, Jason G. McCoy^{1,2}, Patrick F. Chinnery⁴, Konrad J. Karczewski^{2,3}, Sarah E. Calvo^{1,2}, Benjamin M. Neale^{2,3*}, Vamsi K. Mootha^{1,2,5*}

¹ Howard Hughes Medical Institute and Department of Molecular Biology, Massachusetts General Hospital, United States

² Broad Institute of MIT and Harvard, United States

³ Analytic and Translational Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, United States

⁴ Department of Clinical Neurosciences & MRC Mitochondrial Biology Unit, University of Cambridge, United Kingdom

⁵ Department of Systems Biology, Harvard Medical School, United States

* Co-corresponding authors

32 **Abstract**

33 Human mitochondria contain a high copy number, maternally transmitted genome (mtDNA) that
34 encodes 13 proteins required for oxidative phosphorylation. Heteroplasmy arises when multiple
35 mtDNA variants co-exist in an individual and can exhibit complex dynamics in disease and in
36 aging. As all proteins involved in mtDNA replication and maintenance are nuclear-encoded,
37 heteroplasmy levels can, in principle, be under nuclear genetic control, however this has never
38 been shown in humans. Here, we develop algorithms to quantify mtDNA copy number (mtCN)
39 and heteroplasmy levels using blood-derived whole genome sequences from 274,832 individuals
40 of diverse ancestry and perform GWAS to identify nuclear loci controlling these traits. After
41 careful correction for blood cell composition, we observe that mtCN declines linearly with age
42 and is associated with 92 independent nuclear genetic loci. We find that nearly every individual
43 carries heteroplasmic variants that obey two key patterns: (1) heteroplasmic single nucleotide
44 variants are somatic mutations that accumulate sharply after age 70, while (2) heteroplasmic
45 indels are maternally transmitted as mtDNA mixtures with resulting levels influenced by 42
46 independent nuclear loci involved in mtDNA replication, maintenance, and novel pathways.
47 These nuclear loci do not appear to act by mtDNA mutagenesis, but rather, likely act by conferring
48 a replicative advantage to specific mtDNA molecules. As an illustrative example, the most
49 common heteroplasmy we identify is a length variant carried by >50% of humans at position
50 m.302 within a G-quadruplex known to serve as a replication switch. We find that this
51 heteroplasmic variant exerts *cis*-acting genetic control over mtDNA abundance and is itself under
52 *trans*-acting genetic control of nuclear loci encoding protein components of this regulatory
53 switch. Our study showcases how nuclear haplotype can privilege the replication of specific
54 mtDNA molecules to shape mtCN and heteroplasmy dynamics in the human population.

55

56

57 INTRODUCTION

58

59 Mitochondria are ancient organelles that contain a tiny, high copy number circular genome
60 (mitochondrial DNA, mtDNA). Sequencing of the human mtDNA in 1981 (Anderson et al., 1981)
61 revealed that it encodes 13 core protein components of the oxidative phosphorylation system,
62 as well 2 rRNAs and 22 tRNAs required for their expression. The remaining ~1100 mitochondrial
63 proteins, including all proteins required for mtDNA maintenance, replication, and transcription,
64 are encoded by the nuclear DNA (nucDNA) and imported (Rath et al., 2020). Tissues can contain
65 tens to thousands of copies of mtDNA per cell depending on cell type (D’Erchia et al., 2015).
66 Variants in mtDNA can be maternally transmitted or arise somatically, and when they co-exist
67 with wild-type molecules, lead to a state called heteroplasmy. While mtDNA maintenance is fully
68 reliant on nucDNA-encoded proteins, a systematic understanding of how nuclear genetic
69 variation influences variation in mtDNA abundance and heteroplasmy levels in humans is lacking.

70

71 Defects in mtDNA are associated with a spectrum of human diseases (Frazier et al., 2019). Since
72 the first identification of pathogenic mtDNA mutations (Holt et al., 1988; Wallace et al., 1988),
73 scores of maternally inherited syndromes have since been characterized (Ratnaike et al., 2021).
74 Mendelian forms of mitochondrial disease producing mtDNA deletion or depletion were later
75 identified and mapped to nuclear genes involved in mtDNA replication, maintenance, and
76 nucleotide balance (Nishino et al., 1999; Suomalainen et al., 1995; van Goethem et al., 2001).
77 More generally, a quantitative decline in mtDNA copy number (mtCN) and an accumulation of
78 somatic mtDNA mutations have both long been associated with aging and age-associated disease
79 (Ashar et al., 2017; Fazzini et al., 2021; Wanagat et al., 2001). Mutations in mtDNA accumulate in
80 many cancers and in a small subset of tumors fulfill criteria as “drivers” of tumorigenesis (Gopal,
81 Calvo, et al., 2018; Gopal, Kübler, et al., 2018).

82

83 Heteroplasmy dynamics are complex and presumed to be shaped by mutation, drift, and
84 selection. The mtDNA mutation rate has been reported as 10-100x higher than the nucDNA (W.
85 M. Brown et al., 1979; Thomas & Wilson, 1991), with the main non-coding region (control region,
86 CR) containing three hypervariable regions thought to be mutational hotspots (Stoneking, 2000).
87 The high copy number, elevated substitution rate, and lack of recombination have made mtDNA
88 CR variants a valuable genetic tool in anthropology and forensics, even leading to the African
89 mitochondrial “eve” hypothesis (Cann et al., 1987; Vigilant et al., 1991). Heteroplasmy can vary
90 across siblings, attributed to germline bottleneck effects, and between cell types and tissues,
91 thought to be due to random segregation and selection (Li et al., 2015; Walker et al., 2020).
92 Mechanisms underlying heteroplasmy dynamics in humans remain obscure, though classical
93 mouse studies identified nuclear quantitative trait loci (QTLs) controlling mtDNA segregation
94 (Battersby et al., 2003), suggesting a role for nucDNA variation.

95

96 Here, we characterize the spectrum of mtCN and heteroplasmy across ~300,000 individuals
97 spanning 6 ancestry groups in UK Biobank (UKB) and AllofUs (AoU) and identify their nuclear
98 genetic correlates. To our knowledge, this is the largest analysis of human mtDNA sequence to
99 date. After rigorous blood cell composition corrections, we find that mtCN declines with age, is
100 influenced by numerous nuclear genetic loci, and does not decline in most common diseases.

101 mtDNA heteroplasmy shows two patterns: heteroplasmic single nucleotide variants (SNVs),
102 which tend to be somatic and accumulate with age; and heteroplasmic indels, which are more
103 common than SNVs, occur most frequently in the non-coding region, do not vary with age, and
104 are quantitatively inherited as mixtures of multiple alleles along the maternal lineage. These
105 indels are present in most individuals, showing variation across the population and even across
106 single cells from one person. For the first time, we find that many heteroplasmies are influenced
107 by a shared nuclear genetic architecture nominating genes with established roles in mtDNA
108 replication and maintenance as well as mitochondrial genes with no prior links to mtDNA biology.
109 These loci are likely acting by conferring a replicative advantage to specific mtDNA sequences.
110 For instance, the most common heteroplasmy, found in more than 50% of the population, is a
111 length variant in the mtDNA CR, which controls mtCN (in *cis*) and itself is influenced by nuclear
112 loci (*trans*) implicated in a mitochondrial transcription/replication switch.

113

114 RESULTS

115

116 Calling mtDNA copy number and variants at scale

117 We developed mtSwirl, a scalable pipeline for calling mtDNA variants and copy number from
118 whole genome sequencing data (**Methods, Supplementary note 1**). We augmented a pipeline
119 previously used to analyze mtDNA variation in gnomAD (Laricchia et al., 2022), now constructing
120 self-reference sequences for each sample using homoplasmic and homozygous calls on the
121 mtDNA and reference nucDNA regions of mtDNA origin (NUMTs, **Supplementary figure 1A**).
122 mtSwirl shows improved mtDNA coverage, particularly among African haplogroups
123 (**Supplementary figure 1B-E**), and reduced variant calls at very low heteroplasmy
124 (**Supplementary figure 1F**), indicating reduced ancestry- and NUMT-specific mis-mapping. We
125 observe high concordance of heteroplasmy estimates with the prior method used in gnomAD (R^2
126 = 0.996 for heteroplasmy > 0.05), with homoplasmies showing allele fractions now closer to 1
127 suggesting reduced influence of NUMTs (**Supplementary figure 1G**, Laricchia et al., 2022). We
128 used mtSwirl to quantify mtDNA traits across 274,832 individuals of diverse ancestry across UKB
129 and AoU (**Supplementary figure 2, Supplementary table 1**), generating >7,800,000 mtDNA
130 variant calls across all samples.

131

132 Determinants of mtDNA copy number variation

133 We began by identifying covariates of blood mtDNA copy number ($mtCN_{raw}$) in UKB. Our analysis
134 highlights the strong influence of blood cell composition on $mtCN_{raw}$ ($R^2 \sim 23\%$, **Figure 1A**) as
135 previously reported (Hägg et al., 2020; Hurtado-Roca et al., 2016, **Supplementary figure 3C**). We
136 identified several additional technical covariates including time of day, month of year, and fasting
137 duration ($R^2 \sim 2.5\%$, **Figure 1A, Supplementary figures 3E-3J**). Following adjustment for all
138 identified covariates (**Methods, Supplementary note 2, 3**), we find that corrected mtCN (which
139 we term $mtCN_{corr}$) was unimodal in UKB across 178,134 subjects with an average of 61.66 copies
140 per nuclear genome (**Supplementary figure 3D**). We observed a linear decline in $mtCN_{corr}$ with
141 age (**Figure 1C**) of approximately 2% per decade among both males and females.

142

143 We next assessed the degree to which $mtCN_{corr}$ is under nuclear genetic control. Our GWAS
144 identified 92 linkage disequilibrium (LD)-independent nucDNA association signals across 46 loci

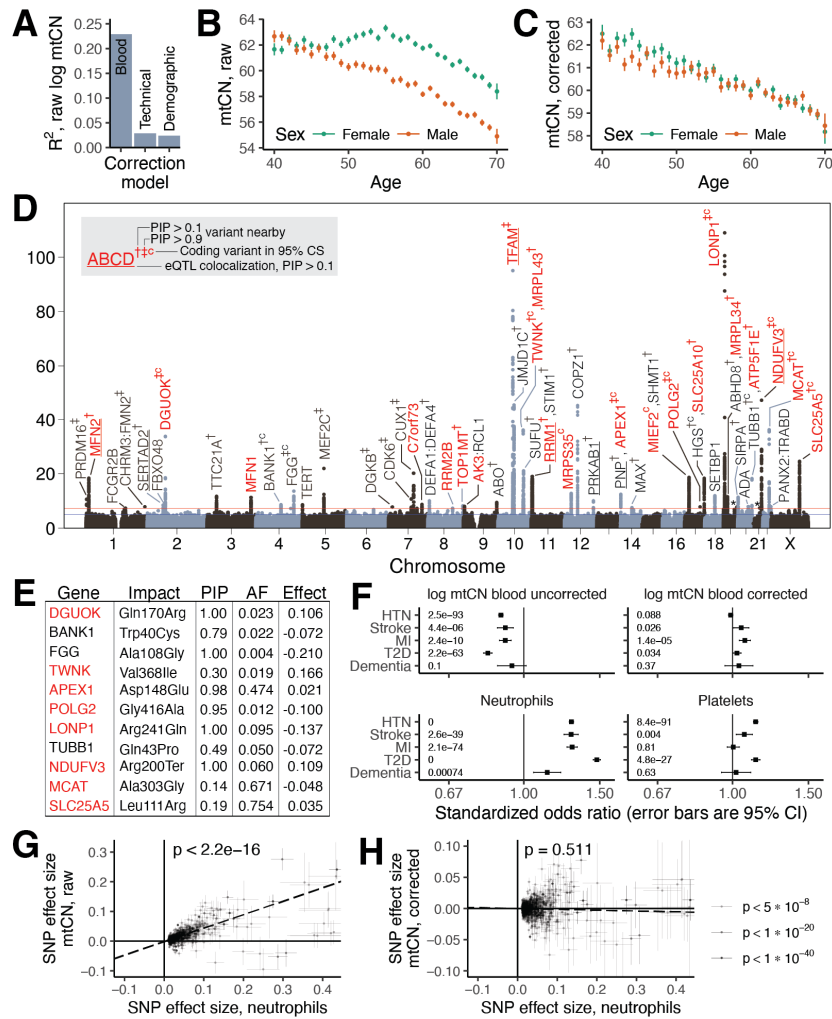


Figure 1. Genetic and phenotypic determinants of mtDNA copy number in UK Biobank. **A.** Variance explained in mtCN by blood composition, technical, and demographic correction models. Relationship of **B.** mtCN_{raw} and **C.** mtCN_{corr} as a function of age and genetic sex. **D.** GWAS Manhattan plot from cross-ancestry meta-analysis in UKB. Labeled genes were obtained either via fine-mapping or, if a credible set (CS) could not be constructed, mapping to the nearest gene. Red genes are mitochondrial or are implicated in mtDNA disease; † = CS variants proximal to the gene with posterior probability of inclusion (PIP) > 0.1; ‡ = CS variants with PIP > 0.9; "c" = coding variant in the CS; underline = eQTL colocalization PIP > 0.1. Asterisks above peaks on chromosome 19 and 21 correspond to GP6 and RUNX1 respectively. **E.** Table of variants in the 95% CS with PIP > 0.1 causing a protein-altering change. Red indicates mitochondria-relevant. **F.** Standardized odds ratios for log mtCN_{raw}, log mtCN_{corr}, and major blood composition phenotypes in predicting risk of selected common diseases in UKB. Inset numbers are p-values; error-bars are 95% CI. HTN = hypertension; MI = myocardial infarction; T2D = type 2 diabetes. Correlation between effect sizes for genome-wide significant lead SNPs detected for neutrophil count between neutrophil count and **G.** mtCN_{raw} and **H.** mtCN_{corr}. Error bars represent 1SE, dotted line is weighted least squares regression line, inset corresponds to regression p-value.

145 (Figure 1D) after cross-ancestry meta-analysis, with an estimated SNP-heritability of ~4%
 146 (Methods). In contrast, mtDNA haplogroup explained < 0.5% of the variance in mtCN_{corr} with only
 147 a few associations of small magnitude observed (Supplementary figure 4A, B). 33 nuclear loci
 148 showed variants with a posterior inclusion probability (PIP) of 0.1 or greater after fine-mapping
 149 (Methods); 11 of these had protein-altering variants in the 95% credible set (CS) at PIP > 0.1

150 (Figure 1E) and seven showed eQTL colocalization with the assigned gene at PIP > 0.1 including
151 *TFAM*, *MFN2*, *NDUFV3*, and *RRM1*. Seven loci contained genes implicated in disorders of mtDNA
152 maintenance, six of which harbored variants with PIP > 0.1. Prioritized genes (Methods) encoded
153 proteins that participate in the mtDNA nucleoid and replisome (*TFAM*, *POLG2*, *TWINKLE*,
154 *TOP1MT*, *LONP1*), nucleotide metabolism (*RRM1*, *RRM2B*, *DGUOK*, *AK3*, *SLC25A5*), and
155 mitochondrial fusion (*MFN1*, *MFN2*). The PNP/APEX1 locus was notable as these adjacent genes
156 encode proteins in nucleotide metabolism and mtDNA repair, neither of which has been
157 implicated in mtCN control. Fine-mapping implicated both genes, even identifying a missense
158 variant in APEX1 at PIP > 0.9 (Supplementary figure 5A). Several additional loci included
159 mitochondrial proteins with no prior links to mtDNA (*SLC25A10*, *MCAT*, *MIEF2*, *NDUFV3*).
160 Telomerase (*TERT*) is in the vicinity of one locus, however fine-mapping did not provide additional
161 evidence for its causality (Supplementary table 3).

162
163 We next tested $mtCN_{corr}$ for heritability enrichment in genes associated with organelles or organs
164 using stratified LD-score regression (S-LDSC, Finucane et al., 2015, 2018; Gupta et al., 2021),
165 Methods). Encouragingly, the most significant organelle enrichment was seen for the
166 mitochondrion (Supplementary figure 4C). Across organs, skeletal muscle and whole blood were
167 top scoring (Supplementary figure 4D). Whole blood enrichment is expected given the sampling
168 site, but skeletal muscle enrichment was unexpected and may be due to shared patterns of gene
169 expression between blood and muscle or non-cell autonomous control of blood mtCN.

170
171 **Blood cell composition confounds prior genetic and phenotypic associations with mtCN**
172 Although many prior studies have reported associations between low blood mtCN and common
173 diseases (Ashar et al., 2017; Chong et al., 2022; Fazzini et al., 2021; Yang et al., 2021), we could
174 not replicate these results using $mtCN_{corr}$ in UKB for type 2 diabetes, myocardial infarction,
175 stroke, hypertension, or dementia (Figure 1F). We tested 24 other common diseases and only
176 observed lower mtCN in individuals with osteoarthritis (Supplementary figure 3K). Upon
177 repeating this using $mtCN_{raw}$, without adjusting for blood composition, we recovered these prior
178 associations (Figure 1F, Supplementary figure 3K). Even the oft-reported elevated mtCN in
179 females (Ding et al., 2015) appears to be largely driven by blood composition (Figure 1B, 1C). Our
180 genetic analyses underscore the confounding effects of blood composition in previous work. We
181 replicated (at $p < 5 \times 10^{-5}$) 70 of the 96 previously reported mtCN GWAS loci (Longchamps et al.,
182 2021) using $mtCN_{corr}$, with 37 at genome-wide significance (GWS) (Methods). However, we
183 recover 12 additional loci from this prior study at GWS using $mtCN_{raw}$ including loci containing
184 *HBS1L-MYOB*, *C2*, *HLA*, *GSDMC*, and *CD226*, which are linked to blood cell types and inflammation
185 (Supplementary figure 4F). In contrast, associations near *TFAM*, a well-known mtCN controlling
186 gene (Ekstrand et al., 2004), strengthen by ~40 orders of magnitude following blood composition
187 correction. It has long been known that inflammation is associated with cardiometabolic disease
188 (Aul et al., 2002); indeed, elevations in inflammatory blood cell indices predict elevated risk for
189 26/29 tested diseases in UKB (Figure 1F, Supplementary figure 3L). Bidirectional Mendelian
190 randomization showed that effect sizes for GWS loci for neutrophil count were strongly positively
191 correlated with corresponding $mtCN_{raw}$ effect sizes (Figure 1G) while the converse did not
192 convincingly hold (Supplementary figure 4G), suggesting that changes in blood cell composition
193 cause $mtCN_{raw}$ changes rather than the reverse. Importantly, neutrophil count effect sizes did not

194 predict corresponding $mtCN_{corr}$ effect sizes (Figure 1H, Supplementary figure 4H). The most
 195 parsimonious explanation for our observations is that previously reported associations between
 196 blood $mtCN$ and common diseases are, in many cases, secondary to blood composition changes.
 197
 198 **Nuclear genetic control of variation in coverage across the mtDNA genome**

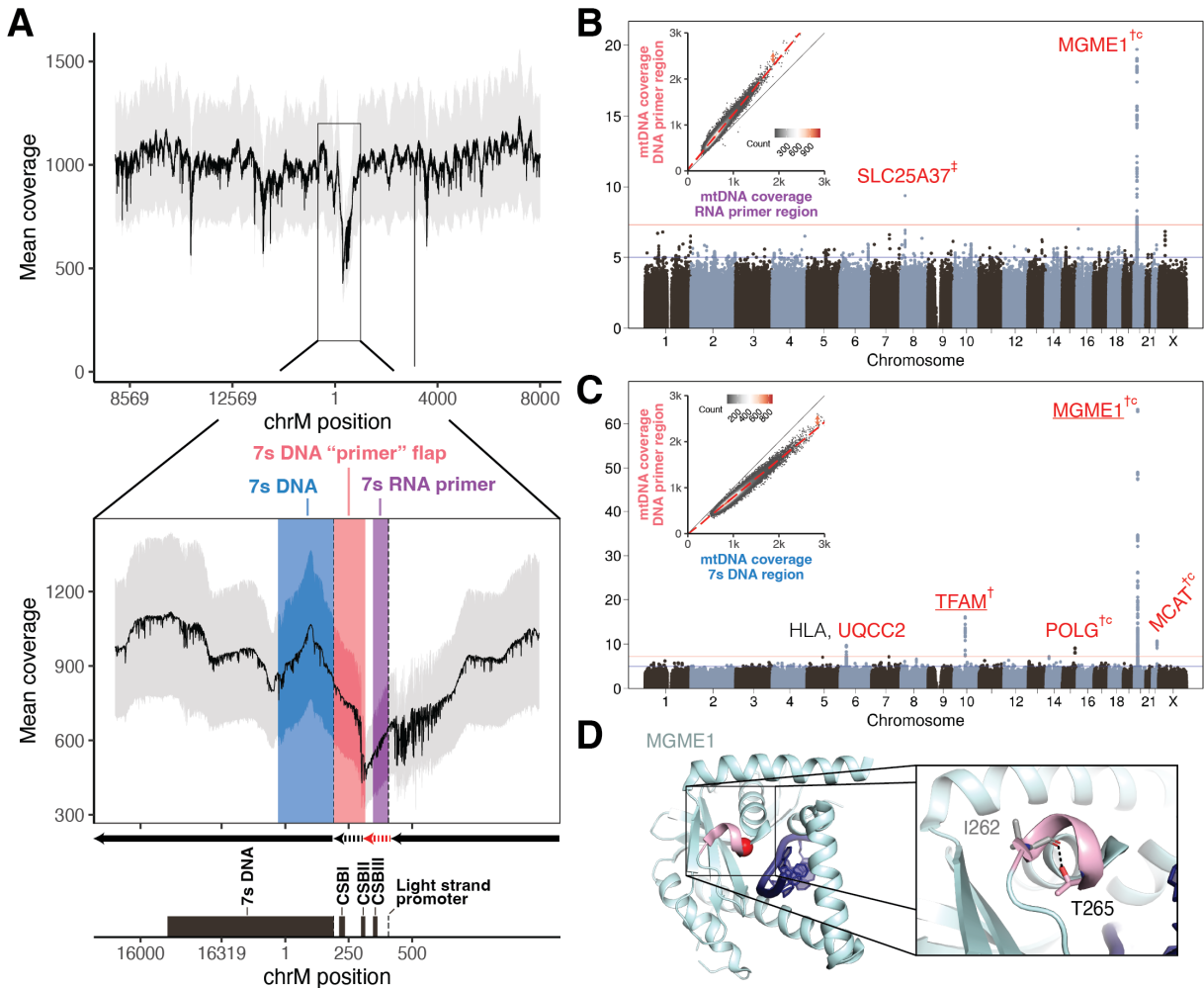


Figure 2. Nuclear genetic control of relative mtDNA coverage in the non-coding region. **A.** Mean per-base coverage across the mtDNA in UKB. Zoomed dropdown highlights coverage depression in the mtDNA non-coding region. Arrows correspond to stages of replication: red dashed arrow = RNA primer; black dashed arrow = transient DNA "primer" flap; black solid arrow = retained replicated DNA. Grey ribbon is +/- 1 standard deviation. CSB = conserved sequence box. **B.** GWAS Manhattan plot of the residual of the regression of mtDNA median DNA primer coverage on median RNA primer coverage. **C.** GWAS Manhattan plot of the residual of the regression of mtDNA median DNA primer coverage on median 7s DNA region coverage. Insets for **B** and **C** show 2D histograms of the correlation between the respective quantities across all individuals in UKB. Red genes are mitochondrial or are implicated in mtDNA disease; † corresponds to CS variants proximal to the gene with posterior probability of inclusion (PIP) > 0.1; ‡ corresponds to CS variants with PIP > 0.9, "c" corresponds to a missense variant in the CS; underline corresponds to eQTL colocalization PIP > 0.1. **D.** Structure of MGME1 (5ZYV) shown with bound ssDNA in dark blue, the 3₁₀ helix in pink and the T265 alpha carbon as a red sphere. Inset shows the hydrogen bond between T265 and I262.

199 Whole genome sequencing (WGS) yields high coverage across the 16,569 bases of the mtDNA,
 200 but it is non-uniform (Figure 2A). We observe a coverage dip by over 50% in the major non-coding

201 segment of the mtDNA called the control region (CR), which contains the light strand promoter
 202 (LSP), three conserved sequence blocks (CSBs), the heavy strand origin of replication (O_H), and
 203 the D-loop, which contains a stable third strand of DNA (7s DNA) (**Supplementary figure 6**).
 204 mtDNA replication starts with RNA primer synthesis from LSP-CSBII (red dashed arrow, **Figure**
 205 **2A**). Primed mtDNA synthesis begins at CSBII, with the nascent DNA between CSBII and O_H
 206 forming a transient flap called the “DNA primer” (black dashed arrow, **Figure 2A**). Further
 207 replication produces the persistent 7s DNA after which replication proceeds (black solid arrow,
 208 **Figure 2A**; c.f. Falkenberg & Gustafsson, 2020). In theory, we expect the highest local WGS
 209 coverage in the persistently triple-stranded 7s DNA, lower coverage in the transiently triple-
 210 stranded “DNA primer” region, and lowest coverage in the RNA primer region. This is what we
 211 observe (**Figure 2A**).

212
 213 We hypothesized that genetic variation in nuclear-encoded mtDNA replication machinery might
 214 influence the persistence of replication intermediates in the CR. To quantify these intermediates,
 215 we computed the difference in coverage between these three regions across individuals in UKB
 216 (insets, **Figures 2B** and **2C, Methods**). Upon performing GWAS and cross-ancestry meta-analysis
 217 for these traits, we find that nuclear genetic variants near *MGME1* associate with the degree of
 218 coverage discordance between the RNA primer and the DNA primer (**Figure 2B**), while variants
 219 near *TFAM*, *POLG*, *MCAT*, and *MGME1* associate with the discordance between 7s DNA and the
 220 DNA primer (**Figure 2C**). All four genes encode mitochondrial-localized proteins, and *MGME1* and
 221 *POLG* work in concert to resolve flap intermediates (i.e., the DNA primer) via exonuclease activity
 222 during mtDNA replication (Uhler et al., 2016). Missense variants in *POLG*, *MGME1*, and *MCAT* all
 223 show PIP > 0.1 after fine-mapping, and the highest PIP variant at the *MGME1* locus causes
 224 p.Thr265Ile which disrupts a hydrogen bond within a helix-forming part of the DNA binding
 225 pocket of the *MGME1* exonuclease domain (**Figure 2D**), potentially impacting DNA binding. We
 226 also identify a variant causing p.Ala303Gly in *MCAT*, which has no prior connection to mtDNA
 227 maintenance and encodes a component of mitochondrial type II fatty acid synthase.

228

229

Intermediate disease phenotypes in carriers of pathogenic mtDNA mutations

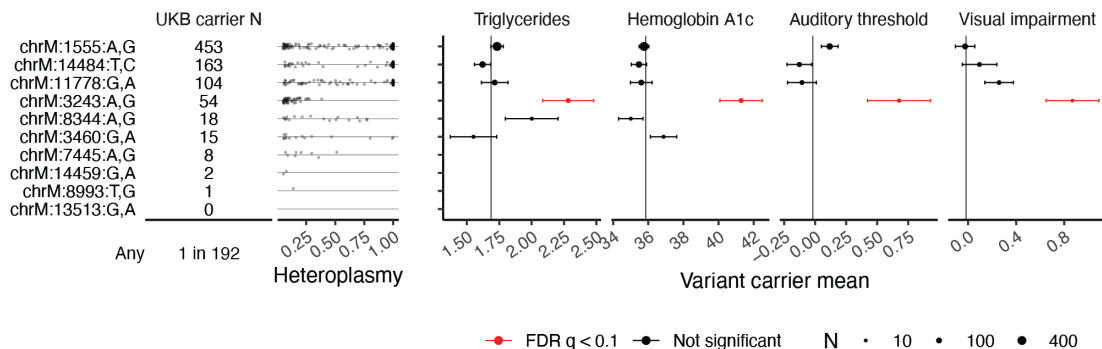


Figure 3. Evidence of intermediate phenotypes among carriers of the MELAS variant in UKB. Table shows carrier frequencies for 10 known pathogenic mutations in UKB, including chrM:3243:A,G (pathogenic for MELAS), with heteroplasmy distributions plotted as jittered points. Panels show mean Hemoglobin A1c, triglyceride levels, auditory threshold (via speech recognition threshold test), and visual impairment (via vision test measured as logMAR) among mtDNA pathogenic variant carriers. Only points corresponding to more than 10 measurements are shown. Vertical lines represent per-trait means among individuals with none of the 10 pathogenic mutations detected.

230 We next considered mtDNA sequence variation in UKB (**Methods**), with an initial focus on ten
231 established pathogenic mtDNA mutations, including those associated with Leber’s hereditary
232 optic neuropathy, mitochondrial encephalopathy, lactic acidosis, and stroke-like episodes
233 (MELAS), and aminoglycoside-induced ototoxicity (**Figure 3**). We find that ~1:192 individuals in
234 UKB carry at least one of the ten variants, in agreement with a previous estimate of 1:200 (Elliott
235 et al., 2008). A longstanding question is whether carriers of rare pathogenic mtDNA variants in
236 the population exhibit intermediate disease phenotypes, which can now be addressed thanks to
237 the rich phenotyping in UKB. We tested four phenotypes traditionally associated with these
238 mtDNA variants: hemoglobin A1c (chrM:3243:A,G), triglyceride levels (chrM:3243:A,G), hearing
239 impairment (chrM:1555:A,G, chrM:3243:A,G, chrM:7445:A,G), and visual impairment
240 (chrM:3460:G,A, chrM:11778:G,A, chrM:14484:T,C, chrM:14459:G,A) (M. D. Brown et al., 2000;
241 Rydzanicz et al., 2011; Sharma et al., 2021; Shoffner et al., 1995). Individuals carrying the
242 chrM:3243:A,G variant show elevated hemoglobin A1c, elevated triglycerides, and hearing and
243 vision impairment (**Figure 3, Methods**). The other tested mtDNA variants were not associated
244 with deviations in these phenotypes.

245

246 **Spectrum of mtDNA sequene variation across 253,583 individuals**

247 Our analysis across UKB and AoU yields the largest database of mtDNA SNVs and indels to date
248 (**Figure 4A**). Consistent with prior gnomAD analysis (Laricchia et al., 2022), we find that the
249 number of homoplasmies per individual is closely related to haplogroup, with haplogroup H
250 (closest to GRCh38 reference) showing the fewest and haplogroup L0 showing the most
251 (**Supplementary figure 7A**). Heteroplasmy distributions were consistent between UKB and AoU
252 (**Figure 4B, Supplementary figure 7D, 7H**), and most individuals carried 0-1 heteroplasmic SNVs
253 and 0-2 heteroplasmic indels (**Supplementary figure 7E**). The hypervariable regions of the
254 mtDNA, found within the non-coding CR, contain an elevated heteroplasmic SNV rate and a vast
255 predominance of heteroplasmic indel variants (**Figure 4A**). Heteroplasmic indels primarily arise
256 near poly-C stretches (e.g., chrM:302, chrM:567, chrM:955, chrM:16182) in the non-protein-
257 coding mtDNA, while coding mtDNA shows a low indel rate despite the presence of many poly-C
258 tracts (**Figure 4A**), consistent with negative selection. We tested the most common
259 heteroplasmies in UKB for associations with risk of 29 common diseases (**Methods**) and found
260 little or no evidence of association, though sample sizes were limited (**Supplementary figure 7F**).

261

262 **mtDNA SNVs and indels exhibit distinct modes of transmission and age accrual**

263 We next investigated the patterns of transmission and age-dependence for mtDNA
264 heteroplasmies. For analysis of age, we focused on AoU given the broader age range of
265 participants (20-90 versus 40-70 for UKB). While heteroplasmic SNVs tend to accumulate with
266 age (particularly after age 70), this was not the case for indel heteroplasmies (**Figure 4C**). Using
267 siblings and parent-offspring pairs in UKB (**Methods**), we find that nearly all heteroplasmic indels
268 are quantitatively maternally transmitted and shared between siblings, while most
269 heteroplasmic SNVs are not (**Figure 4D**). The maternal transmission and stability across age leads
270 us to conclude that most indel heteroplasmies are inherited as mixtures; in contrast, for
271 heteroplasmic SNVs, the typical lack of transmission and accumulation with age strongly suggests
272 that they typically arise via somatic mutagenesis. In contrast to prior reports, no variants showed
273 evidence of paternal transmission (**Figure 4D**). Transitions were far more frequent than

274 transversions and showed a sharp increase in frequency in older age, consistent with the somatic
275 mtDNA mutational spectrum seen in aging brains (Kennedy et al., 2013). Curiously, we observed
276 a decline in transversion heteroplasmies in older individuals (**Supplementary figure 7F**).

277

278 **Nuclear genome GWAS for mtDNA heteroplasmy**

279 We next sought to determine the extent to which mtDNA heteroplasmy is influenced by nuclear
280 genetic loci. To our knowledge, nuclear genetic loci influencing individual mtDNA heteroplasmies
281 in humans have never been reported. Given that most common heteroplasmies showed
282 maternal transmission (**Supplementary figure 7H**), we restricted to individuals carrying each
283 heteroplasmy and performed GWAS with the heteroplasmy level as a quantitative trait (**Figure**
284 **4B, Supplementary figure 7I**).

285

286 We identified 42 LD-independent associations across 39 heteroplasmies after cross-ancestry
287 meta-analysis of our UKB GWAS. Our results revealed a shared nuclear genetic architecture for
288 heteroplasmies across mtDNA sites, with nine of 20 unique nuclear loci associated with >1
289 heteroplasmic variant (**Figure 4E, Supplementary figure 9A**). Cross-mtDNA heterogeneity was
290 also observed: chrM:302:A,AC and chrM:302:A,ACC appeared most associated with loci near
291 *SSBP1*, *TFAM*, *LONP1*, and *MCAT*, while the other heteroplasmies were most strongly associated
292 with loci containing *DGUOK*, *PNP*, and *POLG2*. While many genes implicated in heteroplasmy
293 control were also identified in our mtCN GWAS, others were not (e.g., *TEFM*, *POLRMT*, *MTPAP*,
294 *SSBP1*, *ABHD10*; **Figure 4E**). Many associated loci were near genes with established roles in
295 mtDNA replication and maintenance (**Figure 4F**), with missense variants identified within the 95%
296 credible set in *DGUOK*, *LONP1*, *POLRMT*, *MGME1*, and *POLG2* and eQTL colocalization PIP > 0.1
297 seen for *POLRMT*, *POLG2*, and *TFAM*. Of the novel hits, we highlight a locus containing *C7orf73*
298 (**Figure 4E, Supplementary figure 9F**), which encodes a protein recently linked to Complex IV
299 (Sang et al., 2022), suggesting a moonlighting role for this short protein in mtDNA maintenance.

300

301 Zooming in, we see strong effect sizes from PIP > 0.9 variants in or near genes related to
302 nucleotide metabolism (*PNP*, *DGUOK*) and DNA replication (*POLG2*). The likely causal variant for
303 *PNP* (PIP 1, **Supplementary figure 9G**) is in an intron of *PNP* and colocalizes with a strong negative
304 cross-tissue eQTL (multi-tissue $p \sim 0$; colocalization PIP 1; **Supplementary figure 9H-I**; Aguet et
305 al., 2020) this gene is not yet linked to mtDNA disease but performs an analogous reaction to
306 *TYMP* (an mtDNA disease gene) on purines. The likely causal variant for *DGUOK* (PIP 0.99, **Figure**
307 **4G**) results in a p.Gln170Arg missense change within the kinase domain, potentially disrupting
308 the tertiary structure of the protein as this glutamine side chain participates in a number of
309 hydrogen bonds and stacking interactions (**Figure 4H**). The putative causal variant for *POLG2* (PIP
310 1, **Figure 4I**) results in p.Gly416Ala within a predicted anticodon binding domain. This amino acid
311 is highly conserved (**Supplementary figure 9J**) and the mutation impacts a loop near the *POLG2*
312 homodimer surface (**Figure 4J**). These examples highlight variants impacting proteins and
313 producing a large impact on the levels of specific heteroplasmic mtDNA variants.

314

315 To test if heteroplasmy-associated nuclear loci act via mtDNA mutagenesis, we repeated our
316 GWAS re-coding heteroplasmy traits as “case/control”, where for each mtDNA variant, cases
317 showed detectable heteroplasmy and controls did not. We observed little signal (**Supplementary**

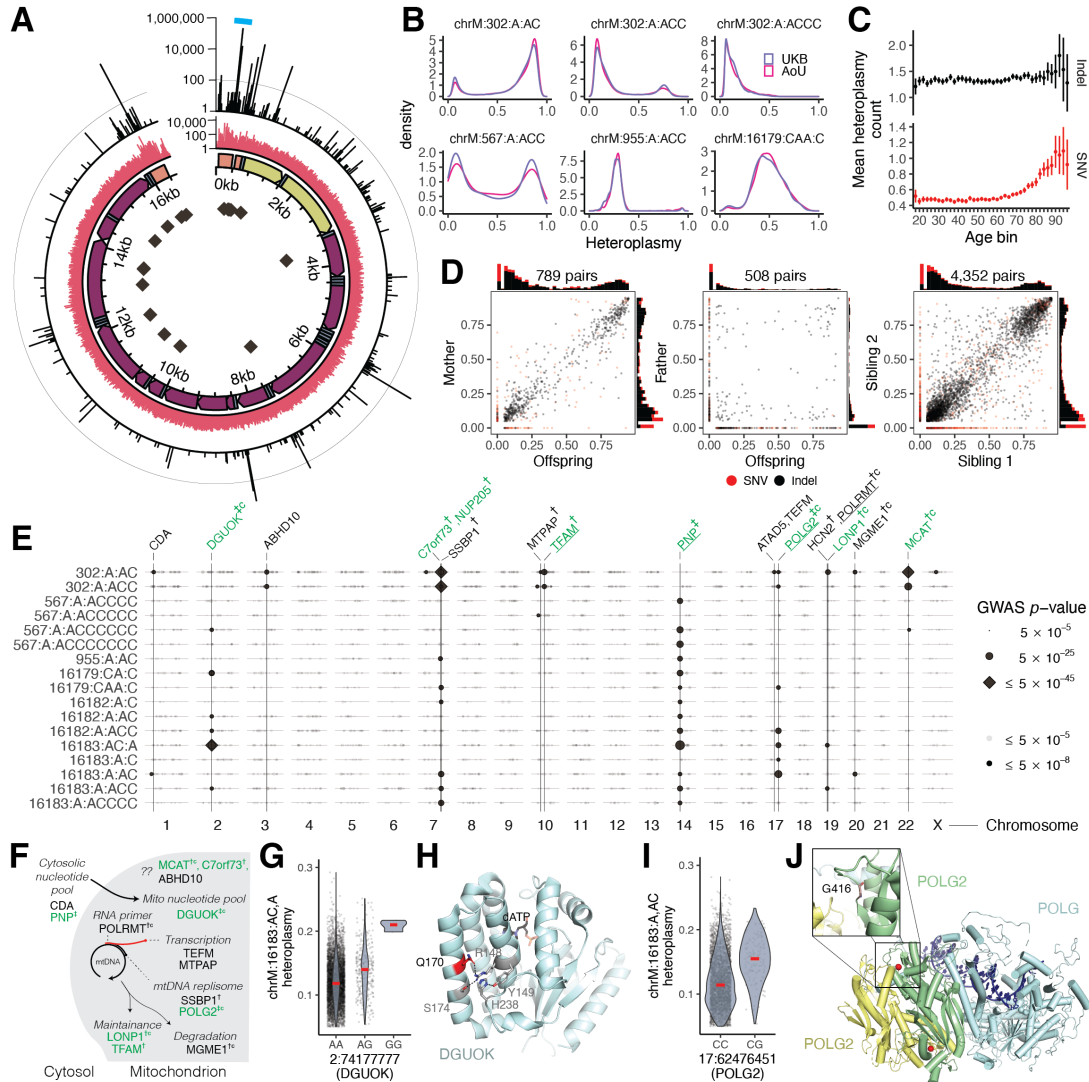


Figure 4. Pervasive nuclear genetic control over the most common mitochondrial DNA heteroplasmies. **A.** mtDNA heteroplasmies passing QC in UKB and AoU. Data tracks show, starting from the inside: positions of poly-C tracts; mtDNA genomic annotations (orange = HVR, yellow = rRNA genes; blue = tRNA genes; purple = coding genes); counts of heteroplasmic SNVs (red); counts of heteroplasmic indels (black). Teal arc corresponds to region highlighted in Figure 5. Light line in outermost track is a reference line at 100. **B.** Selected heteroplasmies distributions across UKB and AoU in individuals carrying the allele. **C.** Mean count of heteroplasmies per individual across age groups in AoU. Error bars are 1SE. **D.** Relationship between heteroplasmies levels in mother-offspring (left), father-offspring (middle), and sibling-sibling (right) for all heteroplasmies found in >5 individuals. **E.** GWAS lead SNPs from all common heteroplasmies with genome-wide significant signals. Point size corresponds to lead SNP p-value; dark points are genome-wide significant. Vertical lines correspond to SNPs near genes of interest and/or loci found across multiple mtDNA variants. Green corresponds to genes nominated for mtCN, + = CS variants with PIP > 0.1; † = CS variants with PIP > 0.9, "c" = coding variant in CS; underline = eQTL colocalization PIP > 0.1. **F.** mtDNA dynamics pathway showing genes highlighted in heteroplasmies GWAS. **G.** chrM:16183:AC,A heteroplasmies as a function of lead SNP genotype in DGUOK. **H.** Structure of DGUOK (2OCP) with amino acid Q170 in red and nearby residues participating in hydrogen bonds or stacking interaction in pink. dATP shown as black sticks. **I.** chrM:16183:A,AC heteroplasmies as a function of lead SNP genotype in POLG2. **J.** Structure of polymerase gamma enzyme (4ZTU) with POLG in light blue and POLG2 subunits in green and yellow. Bound DNA is in dark blue and the POLG2 residue G416 is shown as red spheres. In panels G and I, red lines correspond to medians.

318 **figure 9B)**, arguing against a mutagenic origin influenced by nucDNA variation and supporting the
 319 notion that maternal transmission determines the presence of each tested heteroplasmies, while

320 nuclear variation can influence the subsequent relative heteroplasmic fraction.

321

322 We took several steps to validate our genetic findings. We performed a replication analysis in
323 AoU across 96,698 diverse individuals and observed high concordance between cross-ancestry
324 meta-analysis effect sizes in UKB and AoU ($R^2 = 0.79$, **Supplementary figure 9C**) with limited
325 attenuation (as expected with Winner's curse, c.f. Lohmueller et al., 2003). We investigated
326 potential technical and biological confounders, observing little correlation between these
327 variables and heteroplasmies (**Supplementary figure 8A-E, Supplementary note 2**). We explicitly
328 tested the robustness of our results to the contaminating effects of NUMTs (**Supplementary note**
329 **5**), finding that GWAS effect sizes were not sensitive to mtDNA coverage as would be expected
330 for NUMT-derived signals (**Supplementary figure 8J-M**). Additionally, we found strong
331 correlations between UKB meta-analysis effect sizes and those from individual ancestry groups
332 in AoU despite small N ($R^2 = 0.49-0.78$ **Supplementary figure 9D**), reducing the likelihood of
333 confounding by recent polymorphic NUMTs. We tested all GWAS hits for LD $R^2 > 0.1$ with variants
334 within 20kb windows of 4,736 reference and polymorphic NUMTs, finding only one concerning
335 locus – among UKB EUR, the *SSBP1* locus had LD $R^2 = \sim 1$ with variants in a reference NUMT.
336 Importantly, this locus remained significant for chrM:302:A,AC among AFR in AoU despite AFR
337 showing much lower LD with NUMT variants (**Supplementary figure 9K**).

338

339 **Pervasive length variation in CSBII across individuals and within single cells**

340 The “length heteroplasmy” at chrM:302, located within the CSBII region of the mtDNA CR (**Figure**
341 **5A**), is the most common heteroplasmic site we observed in UKB and occurs within a regulatory
342 motif for mtDNA replication (Wanrooij et al., 2010). Though the reference genome corresponds
343 to G_mAG_7 (nomenclature indicates the length of the poly-G stretch on the GRCh38 opposite
344 strand, **Figure 5A**), we frequently observe individuals harboring G_mAG_8 (chrM:302:A,AC), G_mAG_9
345 (chrM:302:A,ACC), and G_mAG_{10} (chrM:302:A,ACCC). Quantitatively, the levels of these
346 heteroplasmies are shared between siblings (**Figure 5B**), indicating maternal transmission of
347 mixtures of multiple mtDNA haplotypes.

348

349 Most of the 156,885 individuals assessed in UKB harbor a mixture of these length heteroplasmies
350 (**Figure 5C**), with individuals from different haplogroups showing different distributions (**Figure**
351 **5D**). The observed quantitative maternal transmission of heteroplasmy implies that mtDNA
352 mixtures exist in individual cells, and we indeed find mtDNA mixtures at chrM:302 in 171 single
353 cells from one individual (**Figure 5E**) by re-analyzing single-cell mtDNA ATAC-seq data (**Methods**).

354

355 We find multiple lines of evidence linking mtDNA replication and length variation at chrM:302.
356 Longer alleles at this site are associated with declining $mtCN_{corr}$ with an effect size comparable to
357 the *TFAM* locus (**Figure 5F**, PIP ~ 1). GWAS for chrM:302:A,AC, the most common length
358 heteroplasmy, nominated several genes relevant for mtDNA replication and nucleotide balance
359 not identified in other heteroplasmy GWAS (*CDA*, *MTPAP*, *TFAM*, *TEFM*, *LONP1*, *MCAT*; **Figure**
360 **4E, 5G**). $mtCN$ and chrM:302:A,AC heteroplasmy even show colocalization at the two most
361 significant $mtCN$ loci: 10:60145079:A,G (a *TFAM* 5' UTR variant) and 19:5711930:C,T (a *LONP1*
362 missense variant) both show a PIP ~ 1 for $mtCN$ and have PIP > 0.3 for chrM:302:A,AC. It is notable
363 that prior studies have suggested that length variation at the chrM:302 site serves as a “rheostat”

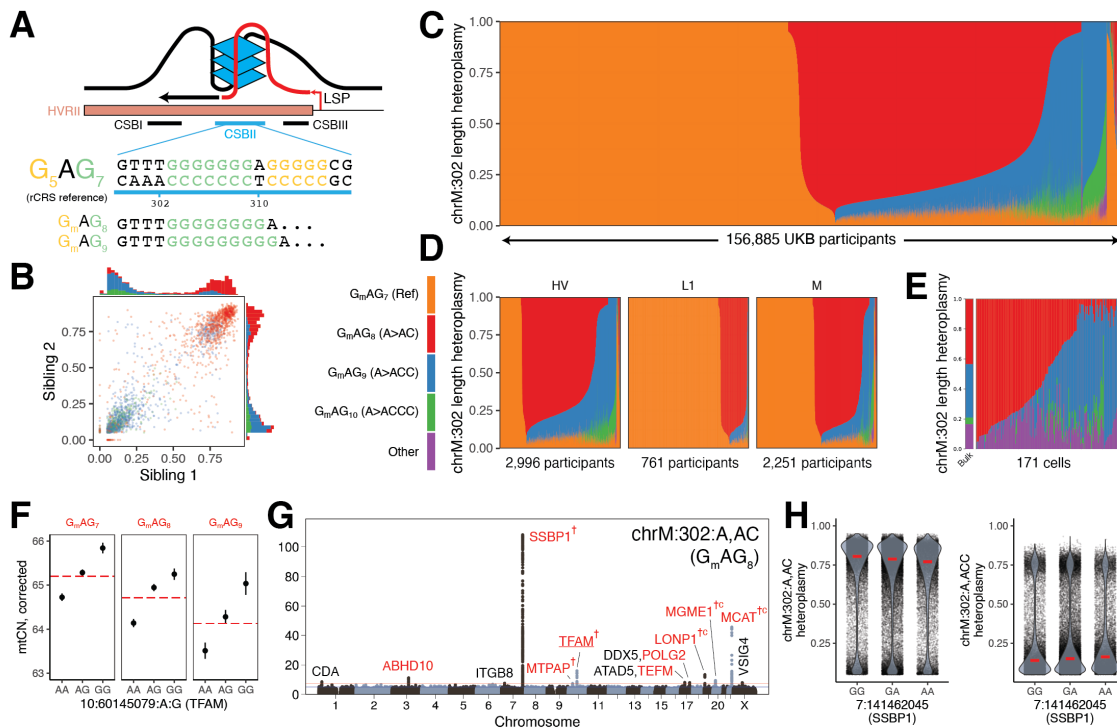


Figure 5. chrM:302 length heteroplasmies are inherited maternally as mixtures, co-exist in single cells, and are under the influence of the nuclear genome. **A.** Scheme showing chrM:302 region inside CSBII responsible for forming a G-quadruplex structure along with length heteroplasmies G_mAG_n nomenclature. **B.** Sibling-sibling transmission of length heteroplasmies at chrM:302. **C.** Length heteroplasmies composition across all UKB individuals. **D.** Length heteroplasmies composition in UKB in select mtDNA haplogroups. **E.** Length heteroplasmies composition across 171 single cells in whole blood. Each vertical bar corresponds to a single individual (C, D) or cell (E). Colors for panels B-E correspond to legend between panels B and D. **F.** Effect of length of major allele at chrM:302 (red line) and TFAM fine-mapped variant (black dot) on mtCN. Error bars are 1SE. **G.** Case-only mtDNA heteroplasmies GWAS Manhattan plot for chrM:302:A,AC. Red genes are mitochondrial or are implicated in mtDNA disease; † corresponds to CS variants proximal to the gene with PIP > 0.1; "c" corresponds to coding variant in CS; underline corresponds to eQTL colocalization PIP > 0.1. **H.** chrM:302 length heteroplasmies as a function of highest PIP SNP genotype in SSBP1 locus. Red line corresponds to per-nuclear-genotype median heteroplasmies.

364 for mtDNA replication versus transcription. The G-quadruplex at CSBII (**Figure 5A**) is a tertiary
 365 structure formed by the DNA and the nascent RNA primer which promotes DNA replication by
 366 blocking RNA polymerase progression (Wanrooij et al., 2010, 2012). *In vitro* studies have
 367 suggested that CSBII G-quadruplex strength is a function of chrM:302 allele, influencing the
 368 degree to which RNA transcription switches to DNA synthesis (**Figure 5A**, Tan et al., 2016). For
 369 the first time we now report that nuclear variants in genes related to the mtDNA replisome can
 370 favor one length heteroplasmies over another – for example, variants near SSBP1 favor
 371 chrM:302:A,ACC (**Figure 5H**). Taken together, our results suggest that nuclear genetic variation
 372 can influence the replication efficiency of mtDNA molecules based on chrM:302 allele.

373

374 DISCUSSION

375

376 mtDNA heteroplasmies dynamics are highly complex, shaped by random drift and selection that in
 377 principle can operate at the level of mtDNA, mitochondria, or cells. Given that all protein

378 machinery for mtDNA replication and maintenance is encoded by the nucDNA, it has long been
379 theorized that the nuclear haplotype could influence mtDNA heteroplasmy. Classical mouse
380 genetics revealed the existence of nuclear QTLs that could influence heteroplasmic mtDNA
381 transmission (Battersby et al., 2003), though specific mechanisms and relevance to humans have
382 been lacking. Here, for the first time, by leveraging whole genome sequencing across two large
383 biobanks, we report pervasive nuclear genetic control of mtDNA abundance and heteroplasmy
384 variation in humans. Many of these nuclear QTLs involve the machinery responsible for mtDNA
385 maintenance, which likely act directly on mtDNA by altering the relative replication efficiency of
386 mtDNA molecules based on their sequence, while several others correspond to genes never
387 before linked to mtDNA biology. High statistical resolution allows us to gain detailed molecular
388 insights into the mechanisms underlying an entire battery of mito-nuclear interactions, with
389 implications for human disease, physiology, and evolution.

390

391 Our ability to dissect the genetic architecture of mtCN and heteroplasmy was possible both
392 because of the statistical power afforded by the scale of large biobanks and because of careful
393 attention given to technical and biological confounders. We analyzed mtDNA sequences across
394 274,832 individuals of diverse ancestries from two biobanks, generating the largest collection of
395 mtDNA traits to date. We were particularly attentive to the technical challenges of contamination
396 by mtDNA pseudogenes in the nuclear genome (NUMTs, **Supplementary Note 5, 6**). We explicitly
397 tested many potential confounders of mtDNA traits, finding that correction of mtCN for blood
398 cell composition had a profound effect on the observed association landscape. Many previously
399 reported associations between blood mtCN and cardiometabolic traits (Ashar et al., 2017; Fazzini
400 et al., 2021) disappear or reverse direction after adjustment for blood cell composition (**Figure**
401 **1F**). Our corrections reduce and even eliminate GWAS hits near genes suspiciously related to
402 blood cell composition and inflammation (e.g., HLA, HBS1L) seen in recent studies (Longchamps
403 et al., 2021). Our data suggest that, in many cases, an inflammatory state in cardiometabolic
404 disease influences blood cell composition, driving the previously observed decline in mtCN.

405

406 The resulting GWAS of mtCN_{corr} and mtDNA heteroplasmies provide new insights into mtDNA
407 maintenance. The nuclear loci we identify, including those with fine-mapped missense variation
408 (e.g., *MGME1*, *POLG*, *POLG2*, *DGUOK*, *LONP1*), are enriched for roles in the mtDNA nucleoid,
409 mtDNA replication, and nucleotide balance, rather than pathways previously implicated in
410 heteroplasmy maintenance in model organisms such as mitophagy or stress response (Gitschlag
411 et al., 2016; Lin et al., 2016). We show how population-level genetic analysis can produce
412 detailed, mechanistic insights into mtDNA replication: GWAS of the relative mtDNA coverage in
413 the 7S DNA “primer” highlights missense variants in both *MGME1* and *POLG*, whose products
414 have exonuclease activity that can resolve this “flap” intermediate. We observe notable
415 differences in the genetic architecture of mtCN_{corr} and heteroplasmy, producing additional
416 insights: while *TFAM*, *LONP1*, *DGUOK*, and *PNP* are associated with both, the former two
417 (encoding components of the mtDNA nucleoid) were the most significant associations for
418 mtCN_{corr}, while the latter two (involved in nucleotide balance) were among the strongest
419 associations across heteroplasmies. QTLs corresponding to *TWNK* were only identified for
420 mtCN_{corr} while associations near *SSBP1*, *TEFM*, and *POLRMT* were specific to heteroplasmy,
421 suggesting that genetic variation in different mtDNA replication genes can have effects specific

422 to mtCN or heteroplasmy. We identify many loci with no prior links to mtDNA biology, such as
423 *C7orf73*, *MCAT*, *ABHD10*, *NDUFV3*, *CDA*, and *ADA*, proposing new roles for their protein products.
424 The *PNP* gene product represents an excellent candidate gene for unsolved mtDNA deletion
425 syndromes as it performs an analogous function to *TYMP* for purines and is notable for its
426 association with the levels of 13 length heteroplasmy variants at three mtDNA sites.

427
428 A striking finding from our work is that nearly everyone harbors heteroplasmic mtDNA variants
429 obeying two key, previously unappreciated, principles: (i) heteroplasmic SNVs are typically
430 somatic, accrue with age sharply after age 70, and tend to be transitions, while (ii) heteroplasmic
431 indels are found in >60% of individuals, do not accrue with age, and are usually inherited as
432 mixtures within the same maternal lineage. Consistent with prior work (Stoneking, 2000),
433 heteroplasmic SNVs tend to occur more in the mtDNA hypervariable regions, but most
434 heteroplasmies detected here are inherited indels. Most heteroplasmic indels appear to occur
435 next to poly-C stretches in the non-protein coding mtDNA; heteroplasmic indel rates are orders
436 of magnitude lower next to poly-C stretches in coding regions, suggesting negative selection in
437 these regions. Strikingly, for any given common indel, we find that maternal heteroplasmy levels
438 quantitatively predict offspring heteroplasmy levels, suggesting neutral transmission. We show
439 for the first time that these heteroplasmy levels are also under nuclear genetic control, with
440 associated loci enriched for genes involved in mtDNA biology and nucleotide balance. These loci
441 are similar across heteroplasmies at multiple mtDNA sites, suggesting a shared genetic
442 architecture.

443
444 Our identified nuclear QTLs for mtDNA length heteroplasmies could, in principle, operate by one
445 of two mechanisms: (1) the associated nuclear variants are “mutagenic” and impair mtDNA
446 copying fidelity resulting in somatic indels due to slippage in poly-C tracts (Marchington et al.,
447 1997), or (2) these nuclear variants confer an mtDNA replicative advantage to maternally
448 inherited mtDNA molecules carrying certain length variants. Our data favors the latter.
449 Case/control GWAS showed very little signal compared to case-only analysis; in concert with the
450 observed maternal transmission this strongly suggests that the identified nuclear QTLs modify
451 existing indel heteroplasmy levels rather than acting via mutagenesis, likely by altering the
452 replicative efficiency of the mtDNA molecules carrying different alleles. Variants near *POLG2*, but
453 notably not *POLG*, were associated with heteroplasmy; *POLG* is the active subunit of mtDNA
454 polymerase in which mutations produce a “mutator” phenotype (Trifunovic et al., 2004), while
455 *POLG2* is the accessory subunit relevant for processivity (Lim et al., 1999; Longley et al., 2006).

456
457 Our work provides insight into mechanisms by which the nuclear haplotype can confer a
458 replicative advantage to specific mtDNA variants. This is perhaps best illustrated by length
459 heteroplasmy at chrM:302. This heteroplasmy occurs within the G-quadruplex in CSBII in the
460 mtDNA noncoding region, which can induce switching from transcription to replication by
461 blocking transcription progression. Prior *in vitro* studies have shown that the chrM:302 length
462 polymorphism impacts the strength of this G-quadruplex hence modifying the
463 transcription/replication switch (Agaronyan et al., 2015; Tan et al., 2016). We find that mixtures
464 of mtDNA with different chrM:302 length variants are maternally inherited in more than half of
465 the population. Once inherited, we show that chrM:302 alleles influence mtDNA abundance

466 (acting in *cis*), and we find that the resulting heteroplasmy levels are influenced in *trans* by
467 nuclear QTLs (e.g., *SSBP1*, *POLG2*, *TEFM*) whose proteins directly operate this regulatory switch
468 (Tan et al., 2016). In sum, our results suggest that the associated nuclear variants alter chrM:302
469 heteroplasmy by influencing factors that interact with the chrM:302 G-quadruplex, thus
470 privileging the replication of mtDNA templates carrying a particular chrM:302 genotype. Recent
471 experiments in embryonic stem cells led to speculation that CSBII length variants may contribute
472 to mtDNA reversion after mitochondrial replacement therapy (MRT) (Kang et al., 2016) due to
473 replicative advantage of carryover mtDNA from the intending mother – our nuclear genetic
474 association results may provide insight into nuclear genetic control of this reversion.

475

476 An open question is why mtDNA heteroplasmy is so common in humans, and whether a selective
477 advantage preserves this variation and the observed mito-nuclear interactions. As the mtDNA
478 has high mutation rates with little or no recombination, it is prone to the accumulation of
479 disabling mutations that could lead to its “meltdown” via Mueller’s ratchet (Lynch et al., 1993).
480 However, mtDNA mutation followed by heteroplasmy is a requisite step in evolutionary
481 adaptation. The identified nuclear QTLs for mtDNA heteroplasmy may represent mechanisms by
482 which a reservoir of such variation can be tolerated and harnessed.

483

484 **ACKNOWLEDGEMENTS**

485 We thank Anna Kotrys, Tom Barton-Owen, and Carla Winter for thoughtful discussions, Melissa
486 Walker for facilitating access to published data, Kristen Laricchia, Katherine Chao, Grace Tiao,
487 Sinéad Chapman, and Namrata Gupta for help with gnomAD data and pipelines, and Wei Zhou
488 for help with SAIGE. This project was supported in part by grants 5R35GM122455 (VKM) and
489 5F30AG074507 (RG) from the National Institutes of Health. VKM is an Investigator of the Howard
490 Hughes Medical Institute. PFC is a Wellcome Trust Principal Research Fellow (212219/Z/18/Z),
491 and a UK NIHR Senior Investigator, who receives support from the Medical Research Council
492 Mitochondrial Biology Unit (MC_UU_00028/7), the Medical Research Council (MRC)
493 International Centre for Genomic Medicine in Neuromuscular Disease (MR/S005021/1), the
494 Leverhulme Trust (RPG-2018-408), an MRC research grant (MR/S035699/1), an Alzheimer’s
495 Society Project Grant (AS-PG-18b-022). This research was supported by the NIHR Cambridge
496 Biomedical Research Centre (BRC-1215-20014). The views expressed are those of the author(s)
497 and not necessarily those of the NIHR or the Department of Health and Social Care. This research
498 has been conducted using the UK Biobank Resource under Application Number 31063. The All of
499 Us Research Program is supported by the National Institutes of Health, Office of the Director:
500 Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2
501 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2
502 OD026551; 1 OT2 OD026555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN
503 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24 OD023121; The
504 Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163;
505 Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community
506 Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; 1 OT2 OD025276. In addition,
507 the All of Us Research Program would not be possible without the partnership of its participants.

508

509 **COMPETING INTERESTS**

510 VKM is a paid advisor to 5am Ventures and Janssen Pharmaceuticals. BMN is a member of the
511 scientific advisory board at Deep Genomics and Neumora, consultant of the scientific advisory
512 board for Camp4 Therapeutics and consultant for Merck. KJK is a consultant for Vor Biopharma.

513

514 **DATA AVAILABILITY**

515 In terms of data processed or generated as part of this study, we provide genetic association
516 statistics for LD-independent lead SNPs and fine-mapped variants in UKB in addition to
517 colocalization results (**Supplementary tables 2-4**). Full GWAS summary statistics from UKB and
518 AoU will be made available in Zenodo upon peer-review. All GWAS sample sizes for each genetic
519 ancestry group, meta-analysis, and phenotype can be found in **Supplementary table 1**. AoU
520 policy does not currently permit public release of individual-level data due to important ethical
521 and privacy considerations: [https://www.researchallofus.org/wp-content/themes/research-](https://www.researchallofus.org/wp-content/themes/research-hub-wordpress-theme/media/2020/05/AoU_Policy_Data_and_Statistics_Dissemination_508.pdf)
522 [hub-wordpress-](https://www.researchallofus.org/wp-content/themes/research-hub-wordpress-theme/media/2020/05/AoU_Policy_Data_and_Statistics_Dissemination_508.pdf)
523 [theme/media/2020/05/AoU_Policy_Data_and_Statistics_Dissemination_508.pdf](https://www.researchallofus.org/wp-content/themes/research-hub-wordpress-theme/media/2020/05/AoU_Policy_Data_and_Statistics_Dissemination_508.pdf)

524

525 In terms of external data used in this study, we leveraged GWAS summary statistics, and ancestry-
526 specific LD-matrices, and a curated list of 29 common, high-quality disease phenotypes
527 generated as part of the Pan UKBB project (*Pan UKBB Initiative, 2022*), with more information
528 available online (<https://pan.ukbb.broadinstitute.org>). UKB phenotype and whole genome
529 sequencing data can be accessed via the UKB Research Analysis Platform after completing a UKB
530 access application: <https://ukbiobank.dnanexus.com/landing>. AoU phenotype and genotype
531 data can be accessed via access to the Controlled Tier v6 on the AoU researcher workbench:
532 workbench.researchallofus.org. Published mtscATACseq data used for chrM:302 analysis can be
533 obtained via approval from dbGaP. Gene-sets for enrichment analyses can be obtained using
534 COMPARTMENTS (<https://compartments.jensenlab.org>) and MitoCarta 2.0
535 (<https://www.broadinstitute.org/files/shared/metabolism/mitocarta/human.mitocarta2.0.html>
536 [\]](https://www.broadinstitute.org/files/shared/metabolism/mitocarta/human.mitocarta2.0.html) as described previously (Gupta et al., 2021). The GRCh37 and GRCh38 reference genomes as
537 well as other standard reference data are available via the GATK resource bundle:
538 <https://gatk.broadinstitute.org/hc/en-us/articles/360035890811-Resource-bundle>. Annotations
539 for the baseline v1.1 and BaselineLD v2.2 models for S-LDSC as well certain other relevant
540 reference data, including the HapMap3 SNP list, can be obtained from
541 <https://alkesgroup.broadinstitute.org/LDSCORE/>. BLASTn was used as available from the NCBI:
542 <https://blast.ncbi.nlm.nih.gov/Blast.cgi>. Known reference and polymorphic NUMTs were
543 obtained from supplemental data as provided in published work (Calabrese et al., 2012; Dayama
544 et al., 2014; Li et al., 2012; Wei et al., 2022).

545

546 **CODE AVAILABILITY**

547 We release the full WDL pipelines for mtDNA analysis from whole genome sequencing data on
548 GitHub: (<https://github.com/rahulg603/mtSwirl>). We also provide the code we used to run the
549 pipeline on the UKB Research Analysis Platform, AoU, and Terra, consolidate all data, and
550 perform mtDNA sample and variant QC. See **Methods** and the README for more information on
551 how to use the pipeline. Several tools were used as part of mtSwirl, including GATK v4.2.6.0
552 (<https://gatk.broadinstitute.org/>), samtools v1.9 (<https://github.com/samtools/samtools>) and
553 bcftools v1.16 (<https://github.com/samtools/bcftools>), Haplochecker 0124

554 (<https://github.com/genepi/haplocheck>), R (r-project.org), Hail (hail.is), and UCSC kent LiftOver
555 tools (genome-source.soe.ucsc.edu/kent.git).

556

557 We used several published tools and scripts to perform downstream analysis of the mtDNA
558 callset in this study. All data wrangling, statistical analysis, and figure generation was performed
559 using either Hail v0.2.98 (hail.is) or R v4.2.1 (r-project.org). Parallelization of tasks in UKB was
560 performed using Hail Batch (batch.hail.is) and in AoU using Cromwell v77
561 (cromwell.readthedocs.io). GWAS was performed in UKB using SAIGE v1.1.5 (saigegit.github.io).
562 For scaling of UKB GWAS, a custom modification of the GWAS pipeline from the Pan UKBB pan-
563 ancestry GWAS was used (https://github.com/atgu/ukbb_pan_ancestry). GWAS was performed
564 in AoU using Hail. mtDNA PCA was performed in R using the irlba v2.3.5.1 package ([https://cran.r-
565 project.org/web/packages/irlba/index.html](https://cran.r-project.org/web/packages/irlba/index.html)). Multinomial models were trained using the nnet
566 v7.3-17 package in R (<https://cran.r-project.org/web/packages/nnet/index.html>). Circos plots
567 were made using the circlize package v0.4.15 in R
568 (https://jokergoo.github.io/circlize_book/book/). For analysis of chrM:302 in single cell data, we
569 used BedTools v2.29.2 (bedtools.readthedocs.io). LD clumping was performed using Plink v1.90
570 (<https://www.cog-genomics.org/plink/>). Finemapping was performed using FINEMAP-inf and
571 SuSiE-inf (<https://github.com/FinucaneLab/fine-mapping-inf>). eQTL data was obtained from
572 GTEx v8 (gtexportal.org) and the eQTL catalogue release 4 (<https://www.ebi.ac.uk/eQTL/>). For
573 replication analysis effect size comparisons, the deming package v1.4 was used in R
574 (<https://cran.r-project.org/web/packages/deming/index.html>). Heritability estimates and
575 enrichment analyses were performed using stratified LD-score regression
576 (<https://github.com/bulik/ldsc>).

577

578

579 **METHODS**

580

581 **Overview of mtSwirl:**

582

583 Here we develop mtSwirl, a scalable pipeline for mitochondrial DNA copy number and variant
584 calling which makes calls relative to an internally generated per-sample consensus sequence
585 before mapping all calls back to GRCh38. In addition to GRCh38 reference files and whole-
586 genome sequencing (WGS) data, the mtSwirl pipeline takes as input nuclear genome reference
587 intervals that represent regions with high homology to the mtDNA (reference NUMTs). We
588 constructed a set of 385 putative NUMTs by using a BLAST-based inventory of reference NUMTs
589 published previously (Li et al., 2012), extending the boundaries of each interval by 500 bases, and
590 merging any overlapping intervals. Initial variant calls within the mtDNA and reference NUMT
591 regions are made from mapped WGS data using Mutect2 and HaplotypeCaller respectively (via
592 GATK v4.2.6.0), and haplogroup inference is performed via Haplogrep (Weissensteiner et al.,
593 2016). Consensus sequences are subsequently constructed using homoplasmies (mtDNA) and
594 homozygous alternate (nucDNA) calls. Reads are realigned to the new consensus sequence and
595 variants are called on the mtDNA using Mutect2. To avoid the artificial coverage depression at
596 the ends of the mtDNA reference genome, we call variants in the control region after alignment
597 to a shifted mtDNA molecule. All variant calls and per-base coverage estimates are then returned
598 to GRCh38 coordinates and output from the pipeline. See **Supplementary note 1** for more details.
599 We release two versions of our pipeline on GitHub (<https://github.com/rahulg603/mtSwirl>):
600 mtSwirlSingle, a single-sample pipeline intended for use with Cromwell and on platforms with
601 high worker limits like Terra and the AllofUs Workbench, and mtSwirlMulti, a multi-sample
602 version which processes multiple samples serially per machine intended for use on platforms
603 with a smaller parallel worker limit such as the UKB Research Analysis Platform (RAP).

604

605 **Cohorts:**

606

607 *UK Biobank (UKB)*

608

609 The UK Biobank is a large prospective cohort study of ~500,000 individuals in the UK (Sudlow et
610 al., 2015), ~200,000 of whom had whole genome sequencing performed at the time of this study.
611 Samples were selected for the first round of WGS using a pseudorandom approach to ensure that
612 included samples were representative of the full cohort. Sequencing data was generated using
613 DNA extracted from buffy coat obtained from participants; more details have been reported
614 previously (Halldorsson et al., 2022). All UKB data was accessed under application 31063 and
615 mtDNA variant calling was performed on the UKB RAP.

616

617 *AllofUs (AoU)*

618

619 AllofUs is a large longitudinal cohort study based in the United States, with a central goal of
620 enrolling a diverse cohort of participants providing electronic health record data over time,
621 specimens for genetic analysis, survey responses, and standardized biometric measurements
622 (“The ‘All of Us’ Research Program,” 2019). At the time of this study, 98,590 individuals had

623 completed whole genome sequencing on samples obtained from whole blood. DNA extraction
624 was completed at the Mayo Clinic, and sequencing was performed at three sequencing centers
625 (Baylor College of Medicine, Broad Institute, and University of Washington) using harmonized
626 protocols. Post-sequencing variant and sample quality control was performed by the AllofUs Data
627 and Research Center (DRC). All mtDNA analyses were performed using the AllofUs Researcher
628 Workbench in the Controlled Tier v6 workspace: “Genetic determinants of mitochondrial DNA
629 phenotypes” using data from the Q2 2022 release. See
630 [https://support.researchallofus.org/hc/en-](https://support.researchallofus.org/hc/en-us/article_attachments/7237425684244/All_of_Us_Q2_2022_Release_Genomic_Quality_Report.pdf)
631 [us/article_attachments/7237425684244/All_of_Us_Q2_2022_Release_Genomic_Quality_Repo-](https://support.researchallofus.org/hc/en-us/article_attachments/7237425684244/All_of_Us_Q2_2022_Release_Genomic_Quality_Report.pdf)
632 [rt.pdf](https://support.researchallofus.org/hc/en-us/article_attachments/7237425684244/All_of_Us_Q2_2022_Release_Genomic_Quality_Report.pdf) for more details on genomics QC and pre-processing.

633
634 *gnomAD v3.1 subset*

635
636 gnomAD v3.1 is a database aggregating whole genome sequencing data from 76,156 samples
637 from several experiments and projects around the world, as part of which an mtDNA variant
638 callset was recently produced (Laricchia et al., 2022). Samples were sourced from several study
639 designs including case-control studies for common diseases, population-based cohorts, and
640 observational studies. Individuals with inborn severe pediatric disease were excluded. Most data
641 are sourced from sequencing performed on either blood samples extracted using study-specific
642 methodologies or from cell lines (Laricchia et al., 2022). We made use of a subset of the gnomAD
643 v3.1 samples to prototype our pipeline (mtSwirl) and compare its performance to previous
644 mtDNA copy number and variant calls (“Vanilla”). We excluded samples with very high mtDNA
645 copy number as done previously (Laricchia et al., 2022) as these are likely cell line samples and
646 not from whole blood; we used a more stringent threshold of 350 as we wanted to maximally
647 enrich for whole blood samples for this trial. We also removed samples with mtCN < 50 due to
648 elevated NUMT contamination in these samples (Laricchia et al., 2022, **Supplementary figure**
649 **7C**). We selected ~6300 samples from gnomAD v3.1 to maximize inclusion of diverse haplogroups
650 including those underrepresented in UK Biobank (**Supplementary figure 2A**). We specifically
651 supplemented samples belonging to the L haplogroups and enforced a cap on the number of
652 samples assigned to either NFE (Non-Finnish European) or FIN (Finnish). For other larger
653 haplogroups we performed random subsampling proportional to the original composition of the
654 gnomAD dataset to achieve our final sample size. All analyses were performed using Terra (*Terra*,
655 n.d.).

656
657 **Computing mean nuclear DNA coverage in UKB:**

658
659 As mean nuclear DNA coverage was not available in UK Biobank, we used samtools v1.9 idxstats
660 (Danecek et al., 2021), samtools flagstat, and GATK v4.2.6.0 CollectQualityYieldMetrics as part of
661 the mtSwirlMulti pipeline to efficiently and economically estimate mean coverage on the nuclear
662 DNA. idxstats-based counts of total mapped reads were computed over autosomes with the
663 subsequent formula applied to get average nuclear DNA coverage after removing contributions
664 from duplicate reads:

665

666 *mean coverage*
667
$$= \frac{(total\ mapped\ reads - singletons - reads\ w/\ discordant\ mate - duplicates) * read\ length}{genome\ length}$$

668

669 **Computing mtDNA copy number:**

670

671 Across all cohorts we use the following formula to compute mtDNA copy number:

672

673
$$2 * mean\ or\ median\ mtDNA\ coverage / mean\ nucDNA\ coverage$$

674

675 We default to use of mean mtDNA coverage for main mtCN-related analyses.

676

677 **Post-calling mtDNA phenotype QC:**

678

679 To integrate our variant calls and perform sample and variant QC, we extended a previously
680 developed pipeline (Laricchia et al., 2022). Single-sample VCFs emitted from mtSwirl were
681 merged into a single Hail MatrixTable (v0.2.98; (Hail Team, n.d.)) upon which all downstream
682 steps were conducted.

683

684 For sample QC, any samples showing homoplasmic variant overlap (see **Supplementary note 1**)
685 were removed. We observed a significant elevation in heteroplasmic SNV calls among samples
686 with mtCN below 50, with a stabilization of heteroplasmic calls above 50 mtDNA copies per cell
687 (**Supplementary figure 7C**), highly suggestive of elevated NUMT contamination in the low copy
688 number samples. Thus, to avoid contamination of our results, all samples with mtCN < 50 were
689 removed. Finally, all samples with evidence of contamination > 2% were removed, as estimated
690 by either (1) mtDNA contamination via Haplocheck 0124 (Weissensteiner et al., 2021) in mtSwirl,
691 (2) nucDNA contamination, or (3) the presence of multiple haplogroup-defining variants at
692 abnormally low allele fraction. Given the small count of samples processed in 2006 and
693 abnormally elevated mtDNA copy number estimates in these samples (**Supplementary figure**
694 **3E**), we excluded these samples from all UKB analyses.

695

696 For variant QC, (1) variants with a very low heteroplasmy (< 0.01) were called as reference with
697 a heteroplasmy of 0, (2) variants with heteroplasmy below 0.05 were flagged and removed as
698 these are at high risk of being enriched for NUMT-derived signals, and (3) all variant calls flagged
699 by Mutect2 were removed. For all sites, a minimum coverage threshold of 100 was used to
700 distinguish between homoplasmic reference calls and sites without variant calls due to low
701 variant-calling confidence as done previously (Laricchia et al., 2022). mtDNA variants were
702 annotated using the Variant Effect Predictor (VEP) v101 (McLaren et al., 2016) and dbSNP v151
703 (Sherry et al., 1999). Variants with at least 0.1% of samples passing filters showing a heteroplasmy
704 between 0 and 0.5 were annotated as “common low heteroplasmy”. Variant calls failing QC were
705 coded with a missing heteroplasmy.

706

707 For mtCN, we remove the samples identified during variant callset sample QC showing signs of
708 contamination, abnormal overlapping homoplasmy calls, or which were processed in 2006. Since

709 we expect mtDNA-wide coverage measures, such as mtCN, to be robust to NUMTs, we do not
710 enforce hard cutoffs on mtCN measurements.

711

712 **Construction of mtDNA heteroplasmy phenotypes:**

713

714 We defined our set of common heteroplasmies in UKB as “common low heteroplasmy” variants
715 **(Methods)** which are present as heteroplasmies in at least 500 individuals, resulting in 39
716 variants. We produced two main sets of phenotypes: (1) a “case-only” dataset consisting of
717 heteroplasmy values for these variants where any individuals without the variant detected were
718 coded as missing and (2) as “case-control” dataset where cases consisted of those with any
719 detectable heteroplasmy and controls consisted of those with the variant not detected. In both
720 phenotype schemes, samples identified as homoplasmic for each variant were always coded as
721 missing. For the case-control dataset, only samples which could be accurately inferred as
722 reference for each variant were labeled as controls – specifically, the sample was coded as
723 missing for a variant if it had a coverage < 100 at the site or showed the variant call as QC-fail
724 **(Methods)**.

725

726 For sensitivity analyses, we produced several additional case-only heteroplasmy datasets: (1)
727 where any variant calls supported by an alternate allele depth (AD alt) of less than the mean
728 nuclear DNA coverage of the sample were made missing; (2) where heteroplasmy estimates were
729 corrected for the depth of mtDNA coverage at the variant site after realignment; and (3) where
730 length heteroplasmy estimates at chrM:302 were corrected for median coverage at CSBII. All
731 corrections were performed by obtaining residuals from the linear regression of the
732 heteroplasmy onto the covariate for each variant across all samples prior to genetic analysis.

733

734 **mtDNA phenotype covariate correction approach:**

735

736 We investigated time of day of blood draw, fasting time, assessment date, and assessment center
737 as technical covariates for mtDNA traits. As draw time and assessment date are continuous, we
738 used natural splines in the correction model to flexibly model nonlinear relationships between
739 these covariates and the mtDNA phenotype. We used knots placed roughly seasonally to model
740 seasonal variation in mtDNA phenotypes – these corresponded to 3-month increments starting
741 on July 1st 2007 and ending on July 1st 2010. For draw time, we used a natural spline basis with 5
742 degrees of freedom. Assessment month and assessment center were modeled as indicator
743 variables. Fasting times were provided in increments of 1 hour and thus were modeled as
744 indicator variables; fasting times of > 18 hours were labeled as 18 and fasting times of 0 were
745 labeled as 1. All terms were included in a joint model for correction.

746

747 We also investigated the relationship between mtDNA phenotypes and blood cell type
748 percentages and mean blood cell volumes. We selected all non-redundant traits available: white
749 blood cell leukocyte count, haematocrit percentage, platelet crit, monocyte percentage,
750 neutrophil percentage, eosinophil percentage, basophil percentage, reticulocyte percentage,
751 high light scatter reticulocyte percentage, immature reticulocyte fraction, mean corpuscular
752 volume, mean reticulocyte volume, mean sphered cell volume, mean platelet thrombocyte

753 volume. We did not include nucleated red blood cell percentage as only ~1% of the entire UKB
754 cohort has non-zero values for this measure, and we excluded lymphocyte percentage given
755 collinearity with neutrophil percentage ($r = 0.92$) and the sum-to-1 property of the white blood
756 cell (WBC) differential measurements. To avoid excess leverage from outlying blood cell
757 measurements, we removed any blood measurements with a Z-score > 4 . All terms were included
758 in a joint model for correction.

759
760 For both the technical covariate and blood cell type models, F-test p-values were obtained for
761 each of the 40 mtDNA phenotypes (39 case-only heteroplasmies and mtCN). For any phenotypes
762 which showed F-test p-values $< 0.05/40$ (Bonferroni corrected), we produced corrected versions
763 of the phenotype by obtaining the residuals from the regression of the mtDNA phenotype onto
764 covariates of interest prior to genetic analysis. For mtDNA copy number, adjustments were
765 performed with $\log(\text{mtCN})$ as the response variable. For heteroplasmy estimates, adjustments
766 were performed with case-only heteroplasmies as the response variable. The specific corrections
767 implemented were:

768
769 $\log \text{mtCN} \sim ns(\text{blood draw time}, 5) + \text{assessment center} + \text{fasting time}$
770 $+ ns(\text{assessment date}, SEASONAL\ KNOTS) + \text{month of assessment}$
771 $+ \text{blood cell variables}$

772
773 As sensitivity analyses for case-only heteroplasmy phenotypes, residuals from the following
774 models were produced:

775
776 $\text{chrM}: 567: A, ACCCCC \sim ns(\text{blood draw time}, 5) + \text{assessment center} + \text{fasting time}$
777 $+ ns(\text{assessment date}, SEASONAL\ KNOTS) + \text{month of assessment}$

778
779 $(\text{chrM}: 16093: T, C; \text{chrM}: 16182: A, ACC; \text{chrM}: 16183: A, AC) \sim \text{blood cell variables}$

780
781 For each response variable, residuals were generated using `residuals(lm(model))` as
782 implemented in R v4.2.1. In all visualizations of corrected variables (e.g., $\text{mtCN}_{\text{corr}}$), we rescale
783 the residualized variable by adding the pre-corrected mean. In the case of $\text{mtCN}_{\text{corr}}$, we rescale
784 the residualized variable and then exponentiate the same to return corrected values back to an
785 absolute scale. See **Supplementary note 2** and **3** for more details.

786

787 **mtDNA PCA and predictive power for mtDNA haplogroups:**

788

789 To construct a high-quality variant genotype matrix for PCA, we obtained the set of homoplasmic
790 variants (heteroplasmy ≥ 0.95) passing QC identified at a MAF ≥ 0.001 in UKB. Any samples
791 with a QC-pass homoplasmy detected were coded as 1 for each respective variant; all others
792 were coded as 0. This binary genotype matrix was subsequently filtered to the set of unrelated
793 samples upon which we computed the first 50 principal components after centering and scaling
794 using the efficient truncated singular value decomposition algorithm implemented in the `irlba`
795 v2.3.5.1 package in R. Related samples were projected onto these PCs to produce a set of mtDNA-
796 PC coordinates for each sample. The set of related samples were defined previously in the Pan

797 UKBB project (*Pan UKBB Initiative*, 2022). In brief, PC-relate was used as implemented in Hail
798 within each assigned genetic ancestry group in UKB and the maximal set of unrelated samples
799 were identified via the maximal independent set algorithm implemented in Hail.

800

801 To assess the goodness of fit of mtDNA PCs for the prediction of top-level mtDNA haplogroups,
802 we fit a multinomial model with top-level haplogroup as the response variable and the first 30
803 mtDNA PCs as explanatory variables as implemented in the `nnet` v7.3-17 package in R (Venables
804 & Ripley, 2002). We only included samples belonging to haplogroups with at least 30 samples in
805 UKB. For assessment of the predictive power of mtDNA PCs for “level 2” haplogroups, we fit
806 multinomial models using a similar approach within each top-level haplogroup, with “level 2”
807 haplogroups as the response variable. In all cases, a null model was fit in parallel with the same
808 response variable with only an intercept term. We computed McFadden’s pseudo- R^2 for each
809 model via the following formula:

810

$$811 \quad \text{pseudo}R^2 = 1 - \frac{\log \text{likelihood}}{\text{null model log likelihood}}$$

812

813 **Correlations between mtCN, mtCN_{corr}, blood cell composition, heteroplasmies, and disease** 814 **phenotypes**

815

816 We obtained common disease diagnoses from UKB via a previously curated set of phecodes and
817 ICD10 codes corresponding to major common diseases (*Pan UKBB Initiative*, 2022) along with
818 demographic variables (age, sex) and blood cell composition phenotypes (**Methods**). We
819 obtained mtCN_{raw}, mtCN_{corr}, common (N > 500) case-only heteroplasmies (**Methods**), and three
820 major blood cell composition traits (platelet crit, monocyte count, and neutrophil count) and
821 performed z-score transformation for each. To test for associations with disease phenotypes, we
822 used a logistic regression model via the `glm` function in R, including age, sex, age², age²*sex,
823 age*sex, top-level haplogroup, and genetic ancestry group assignment as covariates:

824

$$825 \quad \text{disease phenotype} \sim \text{trait} + \text{age} + \text{sex} + \text{age}^2 + \text{age}^2 * \text{sex} + \text{age} * \text{sex} + \text{pop} \\ 826 \quad + \text{top level haplogroup}$$

827

828 We included haplogroups with at least 30 individuals represented in UKB. Odds ratios were
829 obtained as $\exp(\beta_{\text{trait}})$, and the 95% CI was obtained as $\exp(\beta_{\text{trait}} \pm 1.96 * SE_{\text{trait}})$.

830

831 **Derivation of mtDNA coverage discrepancy phenotypes:**

832

833 We obtained mtDNA intervals corresponding to the 7s DNA, heavy strand origin, CSBII, CSBIII,
834 and the light strand promoter (LSP) (Falah et al., 2017; Tan et al., 2016; Xuan et al., 2006). We
835 computed per-individual median mtDNA coverages within the regions corresponding to the first
836 third of the 7s DNA (“7s DNA”), the region between CSBII and the heavy strand origin (“7s DNA
837 primer”), and the region between CSB III and the LSP (“7s RNA primer”). To generate coverage
838 discrepancy phenotypes, we regressed DNA primer coverage onto either 7s DNA coverage or 7s
839 RNA primer coverage. To avoid coverage discrepancies attributable to inherited mtDNA variation

840 within the regions of interest, we included indicator variables for all top-level haplogroups with
841 at least 30 samples as well as their interactions with 7s DNA or 7s RNA primer coverage. The
842 residuals from the following model were used as the coverage discrepancy phenotype for GWAS:

$$843 \quad 7s \text{ DNA primer coverage} \sim (7s \text{ RNA primer or } 7s \text{ DNA coverage}) + \text{haplogroup} \\ 844 \quad + (7s \text{ RNA primer or } 7s \text{ DNA coverage}) * \text{haplogroup} \\ 845 \\ 846$$

847 **Relatedness analyses in UKB:**

848
849 Relatedness was computed and sibling-sibling and parent-offspring pairs were inferred as
850 previously described in UKB (Karczewski et al., 2022). For the assessment of transmission of all
851 QC-pass mtDNA variants, we restricted to only variants found in 5 or more samples.

852

853 **Determination of chrM:302 length heteroplasmy composition:**

854

855 To construct length heteroplasmy profiles, we obtained all post-QC variant calls made at position
856 chrM:302. We defined a “reference” call at chrM:302 for each sample as $1 -$
857 $\text{sum}(\text{heteroplasmy of any allele at chrM:302})$. All samples without variant calls at
858 chrM:302 were assigned a reference fraction of 1, with samples with a depth of < 100 at chrM:302
859 (after local re-alignment during variant calling) excluded. For each sample, we combined all
860 heteroplasmies from calls other than reference, chrM:302:A,AC, chrM:302:A,ACC, and
861 chrM:302:A,ACCC into an “Other” category. Any calls with a missing value for a chrM:302 allele
862 were imputed as a heteroplasmy of 0 for the purposes of visualizations and analyses.

863

864 **Associations between pathogenic variant carrier status and continuous phenotypes in UKB:**

865

866 We obtained continuous phenotypes available in UKB corresponding to classic symptoms of
867 MELAS – diabetes-like symptoms (elevated triglycerides (ID 30870), elevated hemoglobin A1c (ID
868 30750)) and hearing impairment (via the speech-reception-threshold estimate (IDs 20019 and
869 20021)) – as well as the results from the visual acuity test for analysis of known pathogenic
870 variants for Leber’s hereditary optic neuropathy (logMAR from visual acuity test (IDs 5201 and
871 5208)). All obtained phenotypes were filtered to samples with available mtDNA variant calls and
872 corrections were applied for age, sex, age², age²*sex, age*sex, and genetic ancestry group
873 assignment by obtaining residuals from the following linear regression model using
874 `residuals(lm(model))` in R:

875

$$876 \quad \text{measurement} \sim \text{age} + \text{sex} + \text{age}^2 + \text{age}^2 * \text{sex} + \text{age} * \text{sex} + \text{pop} \\ 877$$

878

879 As blood biomarkers tend to have log-normal distributions, corrections were applied after log
880 transformation of HbA1c and triglyceride levels. Post-correction, all measurements were
881 returned to their original scale by adding the pre-correction dataset-wide means for each
882 measurement modality. Final estimates for the speech-recognition-threshold and vision logMAR
883 were generated by averaging measurements for the left and right ear and eye respectively.

883

884 Carriers of known pathogenic mtDNA variants were defined as individuals carrying the variant
885 post-QC at any fraction. We defined a set of controls as individuals with none of the ten known
886 pathogenic mtDNA variants tested. Only samples which could be accurately inferred as reference
887 for all ten variants were labeled as controls – the sample was excluded if, for any of the ten
888 variants, it had a coverage of below 100 at the site or showed a QC-fail variant call (**Methods**).

889
890 Comparisons between residual phenotype values among variant carriers versus global controls
891 were performed only for variant-phenotype pairs with more than 10 defined phenotype values
892 among variant carriers. P-values were obtained by performing a two-sample t-test between
893 phenotype values among variant carriers and the set of global controls, and q-values were
894 obtained by applying the Benjamini-Hochberg procedure.

895
896 **Creation of mutational spectrum categories:**

897
898 Heteroplasmic SNV mutation types in AllofUs were constructed using the set of QC-pass
899 heteroplasmic SNVs. For each SNV type, the set of individuals without any heteroplasmic variants
900 was identified as those with no QC-pass variant call of that type; these individuals were included
901 as zeros in estimates of the mean SNV count of each type.

902
903 **chrM:302 length heteroplasmy inference in single cells:**

904
905 We used the BedTools (Quinlan & Hall, 2010) intersect tool (v2.29.2) to identify read alignments
906 completely spanning the chrM:300-318 locus in the mtscATAC-seq data from Walker et al., 2020,
907 obtained with Massachusetts General Hospital IRB approval under protocol #2016P001517. We
908 then iterated over these reads and classified their chrM:302 length variant by extracting the poly-
909 C/G tracts using a regular expression, 'AA(CCC+[CT]CC+)GC', anchored on the two constant bp on
910 either side of the variant region to detect the canonical variant structure of two poly-C/G tracts
911 with or without a single intervening A/T. Alleles in matching reads were classified based on the
912 length of their poly-C/G tracts, while alleles in the reads that did not match the regular expression
913 were classified as NA. Next, we filtered out any reads with cell barcodes that were not in the
914 published list of cell calls, and further restricted our analysis to only the cells with at least 20
915 reads at the chrM:300-318 locus. For each of these high coverage cells, we calculated the fraction
916 of reads showing each of the top three most common length variants (G_6AG_8 , G_6AG_9 , and G_6AG_{10})
917 and aggregated any other detected alleles into the remainder (other) for display as a stacked bar
918 plot. We also estimated bulk heteroplasmy by summing the allele counts from the high coverage
919 cells and re-calculating the fractions for the top three length variants, again with all other alleles
920 being aggregated into the remainder "other" category.

921
922 **UKB GWAS approach:**

923
924 All GWAS was performed in UKB using approaches as performed in the Pan UKBB initiative (*Pan*
925 *UKBB Initiative*, 2022). In brief, ancestry assignment was performed by projecting UKB samples
926 into genotype PC-space constructed from reference samples from 1000 Genomes (1KG) phase 3
927 and the Human Genome Diversity Project (HGDP) and subsequently using a random forest

928 classifier to assign continental labels trained on the 1KG+HGDP reference data. Within each
929 ancestry group, PCA was performed among unrelated samples with related samples projected
930 onto this PC-space. Further sample QC was performed excluding samples as described as part of
931 the Pan UKBB initiative (*Pan UKBB Initiative, 2022*), including removal of ancestry outliers using
932 a centroid-based metric, individuals with high genotype missingness, sex discordance, and sex
933 chromosome aneuploidies. Variant QC was also performed on UKB-provided imputed v3 variants
934 as part of the Pan UKBB initiative (*Pan UKBB Initiative, 2022*), including only those with INFO
935 scores > 0.8 on autosomes and the X-chromosome. Association tests were performed only on
936 variants with a minor allele count (MAC) > 20 .

937

938 For GWAS, SAIGE v1.1.5 (Zhou et al., 2018) was used to perform association tests within each
939 assigned ancestry group using the first 10 per-population PCs, age, age*sex, age², and age²*sex
940 as covariates (referred to as “baseline”). Ancestry groups were only included if at least 50
941 individuals had the phenotype defined. The use of the SAIGE GRM-based approach allowed for
942 the inclusion of related samples in the GWAS, and we enabled leave-one-chromosome-out fitting
943 in all steps. For all continuous phenotype GWAS (case-only mtDNA heteroplasmy traits and
944 mtCN), phenotypes were inverse rank normalized prior to genetic analysis.

945

946 For all main mtDNA heteroplasmy analyses, top-level mtDNA haplogroup was included as an
947 additional set of covariates in the GWAS model as a set of 24 indicator variables with haplogroup
948 A as reference. Any samples belonging to top-level haplogroups with fewer than 30 samples
949 represented were excluded. The same GWAS model was used for sensitivity analysis of case-only
950 heteroplasmies after removing calls with AD alt $<$ mean nucDNA coverage, after correction for
951 local variant coverage, after correction for CSBII coverage, and after correction for technical or
952 blood trait covariates (**Methods**). For the main mtCN analyses, we used only the baseline
953 covariates to perform genetic associations with mtCN_{raw} and mtCN_{corr}.

954

955 We performed two additional sensitivity analyses for case-only heteroplasmy GWAS: (1) inclusion
956 of 30 mtDNA PCs as covariates in the GWAS model instead of top-level haplogroup for 7 variants
957 which showed relatively high heterogeneity across level 2 haplogroups, and (2) inclusion of mtCN
958 as a covariate in the GWAS model for all case-only heteroplasmies in addition to top-level
959 haplogroup. We also tested the effects of including top-level haplogroup indicator variables as
960 additional covariates in GWAS for mtCN_{raw} and mtCN_{corr}.

961

962 **AllofUs GWAS approach:**

963

964 We performed GWAS in AllofUs as replication for our main case-only heteroplasmy analyses in
965 UKB. Ancestry inference was performed upstream by the AllofUs Data and Research Center
966 (DRC). In brief, AoU samples were projected into the PCA space of genotypes from chromosomes
967 20 and 21 from HGDP and 1KG, and a random forest classifier trained to identify ancestry labels
968 in 1KG+HGDP was used to assign AoU samples continental ancestry labels.

969

970 We performed sample and variant QC after WGS variant calls were imported into Hail. Multi-
971 allelic sites were split and sites with very low pre-computed AF were removed (MAF > 0.0001

972 retained). For sample QC, samples flagged by the DRC as population outliers for several metrics
973 or identified as related by the DRC were excluded. For variant QC, we removed any variants
974 filtered by the DRC, which occurred in brief because of no high-quality genotypes for the variant
975 (defined as GQ \geq 20, DP \geq 10, AB \geq 0.2 for heterozygotes), excess heterozygotes, or a low
976 quality score for the variant. We further removed any variants not in Hardy-Weinberg equilibrium
977 (one-sided $p \leq 1e-10$) and variants with a call rate \leq 0.95. Finally, we removed any variants with
978 MAC $<$ 20 in each assigned ancestry group.

979

980 We next extracted covariates relevant for our GWAS model. We used a SQL query to obtain date
981 of birth in the controlled data repository and used the provided QC flat files to obtain sex assigned
982 at birth. As date of sample collection was not provided, approximate age was constructed for all
983 analyses by subtracting the year of birth from the year 2021. To address residual stratification
984 within assigned ancestry groups, we produced PCs within each ancestry group using a very similar
985 approach as used in UKB (**Methods**) as we found that the provided PCs did not appropriately
986 handle stratification among positive control phenotypes like height, blood glucose, diastolic
987 blood pressure, and systolic blood pressure (**Supplementary note 4**). We included 20
988 recomputed PCs, in addition to approximate age, age², age*sex, and age²*sex as covariates in
989 the final GWAS model. We did not perform genetic analysis for the MID group as less than 400
990 samples with available WGS data were assigned MID.

991

992 We used Hail with the `hl.linear_regression_rows()` method to perform GWAS after
993 all QC. As described in **Methods**, we performed genetic analysis for all QC-pass case-only mtDNA
994 heteroplasmies with homoplasmic calls set to missing. As this analysis is intended for replication,
995 we included any variants found in 300 or more samples across any ancestry group, resulting in
996 41 variants for genetic analysis. Of these, 36 were also analyzed in UKB; 3 UKB variants were not
997 sufficiently common in AoU for genetic analysis. As in UKB, for the analysis of case-only mtDNA
998 heteroplasmies, top-level mtDNA haplogroup was included as covariates in the GWAS model as
999 a set of 27 indicator variables in addition to age, sex, and PC covariates. Samples belonging to
1000 top-level haplogroups with fewer than 30 samples in AoU were excluded. All case-only mtDNA
1001 heteroplasmy phenotypes were inverse rank normalized prior to analysis.

1002

1003 See the AllofUs genotype quality report for more information on upstream genotype data and
1004 sample QC, ancestry inference, and relatedness inference
1005 ([https://support.researchallofus.org/hc/en-](https://support.researchallofus.org/hc/en-us/article_attachments/7237425684244/All_Of_Us_Q2_2022_Release_Genomic_Quality_Report.pdf)
1006 [us/article_attachments/7237425684244/All_Of_Us_Q2_2022_Release_Genomic_Quality_Repo](https://support.researchallofus.org/hc/en-us/article_attachments/7237425684244/All_Of_Us_Q2_2022_Release_Genomic_Quality_Report.pdf)
1007 [rt.pdf](https://support.researchallofus.org/hc/en-us/article_attachments/7237425684244/All_Of_Us_Q2_2022_Release_Genomic_Quality_Report.pdf)).

1008

1009 **Heritability estimation and enrichment analyses for mtCN:**

1010

1011 Stratified linkage disequilibrium score regression (S-LDSC, Finucane et al., 2015) was used for
1012 heritability estimation and enrichment analyses for mtDNA copy number in UKB as performed
1013 previously (Gupta et al., 2021). In brief, we analyzed EUR summary statistics in UKB, restricting
1014 variants to those in HapMap3 (HM3). We estimated overall SNP-heritability controlling for 97
1015 annotations corresponding to coding regions, enhancer regions, minor allele frequency bins, and

1016 others (Gazal et al., 2017, referred to as baselineLD v2.2). For enrichment analyses, we obtained
1017 gene-sets corresponding to (1) the top 10% of genes specifically expressed in major tissues from
1018 GTEx (Finucane et al., 2018) and (2) genes producing protein products that localize to each major
1019 organelle with high confidence using COMPARTMENTS (Binder et al., 2014). Variants were
1020 mapped to each gene with a 100kb symmetric window and LD scores for each gene-set
1021 annotation were computed using the 1000G EUR reference panel
1022 (<https://alkesgroup.broadinstitute.org/LDSCORE/>). Heritability enrichment for all gene-sets was
1023 tested using S-LDSC atop the baseline v1.1 model, controlling for 53 annotations including coding
1024 regions and 5' and 3' UTRs (Finucane et al., 2015).

1025

1026 **Cross-ancestry meta-analysis in UKB and AllofUs:**

1027

1028 We conducted a fixed-effect meta-analysis across ancestries in each cohort (UKB and AoU) based
1029 on inverse-variance weighted betas and standard errors (de Bakker et al., 2008). For each
1030 ancestry, we excluded low-confidence variants defined as MAC \leq 20 in either biobank. We
1031 computed effect size heterogeneity P-values across ancestries using Cochran's Q-test (Cochran,
1032 1954). All computation was done using Hail v0.2.

1033

1034 All visualizations of main GWAS (e.g., mtCN, coverage discrepancy traits, heteroplasmy traits) are
1035 of cross-ancestry meta-analyses after restriction to the set of "high quality" variants as defined
1036 previously (*Pan UKBB Initiative*, 2022).

1037

1038 **Identification of LD-independent lead SNPs and locus definitions:**

1039

1040 Clumping was performed using Plink v1.90 (Purcell et al., 2007) in Hail Batch for GWAS results
1041 obtained in UK Biobank after filtering to high quality variants. We used significance thresholds of
1042 1 for both the index and clumped SNPs, set the LD threshold for clumping at 0.1, and set the
1043 distance threshold at 500kb. We used single ancestry and multi-ancestry LD reference panels
1044 corresponding to the ancestry groups included in the final multi-ancestry meta-analyses for each
1045 mtDNA phenotype as well as for blood cell traits. Reference panels were constructed by randomly
1046 sampling 5000 individuals from all samples within any given set of ancestry groups in the UK
1047 Biobank. For the single-ancestry LD panels corresponding to ancestry groups with less than 5000
1048 individuals (EAS and MID), the full sample available for each ancestry group was used. More
1049 details on the LD reference panels can be found as part of the Pan UKBB project (*Pan UKBB*
1050 *Initiative*, 2022). Clumping output files from Plink were converted to Hail Tables and then
1051 combined into MatrixTables using the multi-way-zip-join method as implemented in Hail.

1052

1053 We defined distinct loci conservatively by starting with genome-wide significant LD-independent
1054 lead SNPs and merging any SNPs within 2 Mb of one another.

1055

1056 **Replication of previous mtCN GWAS with our study:**

1057

1058 We performed a comparison of significant loci identified in a previous GWAS of mtCN in UKB
1059 (Longchamps et al., 2021) with our own by performing LD clumping on previously released

1060 summary statistics as described (**Methods**) using 1KG phase 3 EUR reference data for LD. We
1061 defined distinct loci as described (**Methods**), merging any SNPs within 2 Mb of one another,
1062 arriving at 96 loci previously identified. We defined a replicated locus with mtCN_{raw} or mtCN_{corr}
1063 as one where our GWAS showed a signal at $p < 5 \cdot 10^{-5}$ or $5 \cdot 10^{-8}$ within 2 Mb of the most
1064 significant variant identified in the previous study within each locus.

1065

1066 **Bidirectional Mendelian randomization between UKB mtCN and neutrophil count:**

1067

1068 GWAS effect sizes and LD-independent loci from the UKB cross-ancestry meta-analysis for raw
1069 mtCN and fully corrected mtCN were obtained. Summary statistics and LD-independent loci from
1070 GWAS among EUR for neutrophil count (ID 30140) were obtained from the Pan UKBB project
1071 (*Pan UKBB Initiative*, 2022). Sites for comparison were restricted to those passing variant QC as
1072 performed in UKB (**Methods**). For each mtCN phenotype, neutrophil count and mtCN GWAS
1073 effect sizes were obtained for all mtCN genome-wide significant variants, and vice-versa, mtCN
1074 and neutrophil count GWAS effect sizes were obtained for all neutrophil count genome-wide
1075 significant variants. We assessed the relationship between pre- and post-correction mtCN GWAS
1076 effect sizes and neutrophil count GWAS effect sizes via inverse-variance weighted linear
1077 regression using weights corresponding to $\frac{1}{SE(mtCN)^2} * \frac{1}{SE(neutrophil)^2}$, where effect size standard
1078 errors were obtained from the respective GWAS.

1079

1080 **Fine-mapping in UKB:**

1081

1082 To identify putative causal variants in associated loci, we conducted statistical fine-mapping of
1083 mtDNA traits in UKB using cross-ancestry meta-analysis summary statistics. While we previously
1084 showed that fine-mapping a meta-analysis is often miscalibrated due to heterogeneous
1085 characteristics of constituent cohorts (e.g., genotyping or imputation) (Kanai et al., 2022), a
1086 within-cohort cross-ancestry meta-analysis like the present study is a notable exception given no
1087 such heterogeneity systematically exists across ancestries.

1088

1089 We used FINEMAP-inf and SuSiE-inf which model infinitesimal effects (Cui et al., 2022), with
1090 cross-ancestry meta-analysis summary statistics (**Methods**) and a covariate-adjusted in-sample
1091 dosage LD matrix (Kanai et al., 2021). We defined fine-mapping regions based on a 3 Mb window
1092 around each lead variant and merged regions if they overlapped as described previously (Kanai
1093 et al., 2021). We excluded the major histocompatibility complex (MHC) region (chr 6: 25–36 Mb)
1094 from analysis due to extensive LD structure in the region. For each method, we allowed up to 10
1095 causal variants per region and derived posterior inclusion probabilities (PIP) of each variant using
1096 a uniform prior probability of causality. To achieve better calibration, we computed min(PIP)
1097 across the methods and derived up to 10 independent 95% credible sets (CS) from SuSiE-inf as
1098 described elsewhere (Kanai et al., 2021). All reported PIP are min(PIP) between the two methods.

1099

1100 **Enrichment of functional categories among fine-mapped variants:**

1101

1102 We computed functional enrichment of fine-mapped variants across the mtDNA traits in UKB.
1103 We first annotated each variant with seven functional categories (pLoF, missense, synonymous,
1104 5' UTR, 3' UTR, promoter, cis-regulatory element [CRE], and non-genic) as described previously
1105 (Kanai et al., 2021). We then estimated functional enrichment for each category as a relative risk
1106 (i.e., a ratio of proportion of variants) between being in an annotation and fine-mapped ($PIP \leq$
1107 0.01 or $PIP > 0.1$). That is, a relative risk = (proportion of variants with $PIP > 0.1$ that are in the
1108 annotation) / (proportion of variants with $PIP \leq 0.01$ that are in the annotation). 95% confidence
1109 intervals are calculated using bootstrapping with 5,000 replicates. We note that, to increase
1110 statistical power, we combined pLoF/missense and 5'/3' UTR into single categories respectively
1111 and used a more lenient threshold ($PIP > 0.1$ vs. > 0.9) compared to our previous analysis (Kanai
1112 et al., 2021).

1113

1114 **Gene- and variant-prioritization:**

1115

1116 To nominate genes using GWAS results, we used the following approach to balance clarity with
1117 confidence in the gene assignment.

1118

- 1119 1. If the locus had a credible set, for each credible set (CS):
 - 1120 a. Filter to variants in the credible set and retain variants from the CS that are either
1121 minimal PIP, coding, or have $PIP > 0.7$
 - 1122 b. If the variant has $PIP > 0.9$ and is a coding variant for a gene, assign that gene to
1123 the CS
 - 1124 c. Otherwise assign genes within 3kb of the variant or, if no genes are within 3kb,
1125 assign the nearest gene to the CS
- 1126 2. If the locus had multiple credible sets and at least one had a variant with $PIP > 0.1$, we
1127 retained assignments only corresponding to variants with $PIP > 0.1$
- 1128 3. If the locus did not have a credible set, we assigned the gene with a boundary nearest to
1129 the most significant variant in the locus

1130

1131 If a variant is inside a gene body (but is non-coding), we consider that gene to be nearest. For
1132 case-only heteroplasmy GWAS, when the same locus was significant across multiple
1133 heteroplasmy phenotypes, we performed manual integration to arrive at a set of genes
1134 supported by the most compelling genetic evidence across variants for each locus. The SSBP1
1135 locus was particularly complex, so we assign SSBP1 (which harbors the max PIP variant) and
1136 provide visualization of the full locus (**Supplementary figure 9K**). We do not use fine-mapping
1137 evidence from variants with $PIP > 0.1$ that are not assigned to a credible set. All assignments were
1138 manually reviewed. In all GWAS visualizations, we label the strength of evidence supporting the
1139 gene assignment (e.g., if supported by moderate or high-PIP fine-mapped variants, coding
1140 variants).

1141

1142 **Colocalization with eQTLs:**

1143

1144 We conducted colocalization of fine-mapped variants of mtDNA phenotypes and *cis*-eQTL
1145 associations from GTEx v8 (Aguet et al., 2020) and eQTL catalogue release 4 (Kerimov et al., 2021)

1146 as described previously (Kanai et al., 2021). Briefly, we retrieved fine-mapping results of *cis*-eQTL
1147 associations that were fine-mapped using SuSiE (Wang et al., 2020) with covariate-adjusted in-
1148 sample dosage LD matrices (Kanai et al., 2021). We then computed a posterior inclusion
1149 probability of colocalization for a variant as a product of PIP for GWAS and for *cis*-eQTL ($CLPP =$
1150 $PIP_{GWAS} \times PIP_{cis-eQTL}$) (Hormozdiari et al., 2016). When displaying colocalization across
1151 heteroplasmy traits, we indicate colocalization if we see colocalization $PIP > 0.1$ for the assigned
1152 gene and any variant in the credible set for any tissue and for any heteroplasmy trait.

1153

1154 **Replication of UKB heteroplasmy results in AllofUs:**

1155

1156 To perform replication analysis in AllofUs, we used LD-independent lead SNPs from all case-only
1157 heteroplasmy GWAS originally performed in UKB (**Methods**). We filtered association statistics
1158 from AoU (**Methods**) to these lead variants and compared effect sizes when the variants were
1159 identified in AoU with $MAC > 20$. We used Deming regression implemented in the `deming` v1.4
1160 package in R to assess the relationship between effect sizes for these lead SNPs in cross-ancestry
1161 meta-analyses in the two biobanks while accounting for standard errors in both (Deming, 1943;
1162 Zhou et al., 2022). We coded alleles such that effect sizes were always positive in UKB.

1163

1164 **Assessment of LD with known polymorphic and reference NUMTs:**

1165

1166 We collated an extensive database of polymorphic and reference NUMT intervals using BLAST,
1167 known reference NUMTs (Calabrese et al., 2012; Li et al., 2012), and published polymorphic
1168 NUMTs inferred using mate-pair mapping discordance (Dayama et al., 2014; Wei et al., 2022). To
1169 search for regions of homology to the mtDNA within the reference nucDNA, we used BLASTn with
1170 the GRCh37 reference genome with a word size of 11, an expect threshold of 0.05, short queries
1171 enabled, and default values for the other parameters. In total, we obtained 4,736 overlapping
1172 reference and polymorphic NUMT intervals. We constructed a 20kb window around each
1173 nucDNA NUMT region (10kb up, 10kb down) and then conservatively tested for $LD R^2 > 0.1$
1174 between any SNP in the window and each lead variant at genome-wide significance for our UKB
1175 case-only heteroplasmy GWAS using in-sample genome-wide EUR LD matrices generated in UKB
1176 (*Pan UKBB Initiative*, 2022). All LD values used to examine individual loci in either biobank was
1177 computed in-sample – for example, in AoU we computed LD using the post-QC genotype
1178 MatrixTable (**Methods**) used for GWAS with the Hail function `hl.ld_matrix()`.

1179

1180 **Multiple alignment of POLG2 protein sequence:**

1181

1182 POLG2 homologs were detected via best bi-directional BlastP hit (Expect $< 1e-3$) from human
1183 and were aligned via MUSCLE (Edgar, 2004).

1184

1185

1186 **REFERENCES**

1187

1188 Agaronyan, K., Morozov, Y. I., Anikin, M., & Temiakov, D. (2015). Replication-transcription
1189 switch in human mitochondria. *Science*, *347*(6221), 548–551.

1190 <https://doi.org/10.1126/SCIENCE.AAA0986>

1191 Aguet, F., Barbeira, A. N., Bonazzola, R., Brown, A., Castel, S. E., Jo, B., Kasela, S., Kim-Hellmuth,
1192 S., Liang, Y., Oliva, M., Flynn, E. D., Parsana, P., Fresard, L., Gamazon, E. R., Hamel, A. R.,

1193 He, Y., Hormozdiari, F., Mohammadi, P., Muñoz-Aguirre, M., ... Volpi, S. (2020). The GTEx

1194 Consortium atlas of genetic regulatory effects across human tissues. *Science*, *369*(6509),

1195 1318–1330. [https://doi.org/10.1126/SCIENCE.AAZ1776/SUPPL_FILE/AAZ1776-](https://doi.org/10.1126/SCIENCE.AAZ1776/SUPPL_FILE/AAZ1776_TABLESS10-)

1196 [S16.XLSX](https://doi.org/10.1126/SCIENCE.AAZ1776/SUPPL_FILE/AAZ1776_TABLESS10-S16.XLSX)

1197 Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I.

1198 C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R., & Young,

1199 I. G. (1981). Sequence and organization of the human mitochondrial genome. *Nature*,

1200 *290*(6), 338–346.

1201 Ashar, F. N., Zhang, Y., Longchamps, R. J., Lane, J., Moes, A., Grove, M. L., Mychaleckyj, J. C.,

1202 Taylor, K. D., Coresh, J., Rotter, J. I., Boerwinkle, E., Pankratz, N., Guallar, E., & Arking, D. E.

1203 (2017). Association of mitochondrial DNA copy number with cardiovascular disease. *JAMA*

1204 *Cardiology*, *2*(11), 1247–1255. <https://doi.org/10.1001/jamacardio.2017.3683>

1205 Aul, P., Idker, M. R., Ifai, A. R., Ynda, L., Ose, R., Ulie, J., Uring, E. B., & Ook, A. R. C. (2002).

1206 Comparison of C-Reactive Protein and Low-Density Lipoprotein Cholesterol Levels in the

1207 Prediction of First Cardiovascular Events. <https://doi.org/10.1056/NEJMoa021993>,

1208 *347*(20), 1557–1565. <https://doi.org/10.1056/NEJMoa021993>

1209 Battersby, B. J., Loredó-Ostí, J. C., & Shoubridge, E. A. (2003). Nuclear genetic control of

1210 mitochondrial DNA segregation. *Nature Genetics*, *33*(2), 183–186.

1211 <https://doi.org/10.1038/ng1073>

1212 Binder, J. X., Pletscher-Frankild, S., Tsafoú, K., Stolte, C., O'Donoghue, S. I., Schneider, R., &

1213 Jensen, L. J. (2014). COMPARTMENTS: Unification and visualization of protein subcellular

1214 localization evidence. *Database*, *2014*, 1–9. <https://doi.org/10.1093/database/bau012>

1215 Brown, M. D., Trounce, I. A., Jun, A. S., Allen, J. C., & Wallace, D. C. (2000). Functional analysis of

1216 lymphoblast and cybrid mitochondria containing the 3460, 11778, or 14484 leber's

1217 hereditary optic neuropathy mitochondrial DNA mutation. *Journal of Biological Chemistry*,

1218 *275*(51), 39831–39836. <https://doi.org/10.1074/jbc.M006476200>

1219 Brown, W. M., George, M., & Wilson, A. C. (1979). Rapid evolution of animal mitochondrial

1220 DNA. *Proceedings of the National Academy of Sciences*, *76*(4), 1967–1971.

1221 <https://doi.org/10.1073/PNAS.76.4.1967>

1222 Calabrese, F. M., Simone, D., & Attimonelli, M. (2012). Primates and mouse NumtS in the UCSC

1223 Genome Browser. *BMC Bioinformatics*, *13*(SUPPL.4), 1–9. <https://doi.org/10.1186/1471->

1224 [2105-13-S4-S15/FIGURES/5](https://doi.org/10.1186/1471-2105-13-S4-S15/FIGURES/5)

1225 Cann, R. L., Stoneking, M., & Wilson, A. C. (1987). Mitochondrial DNA and human evolution.

1226 *Nature*, *325*(6099), 31–36. <https://doi.org/10.1038/325031A0>

1227 Chong, M. R., Narula, S., Morton, R., Judge, C., Akhbar, L., Cawte, N., Pathan, N., Lali, R.,

1228 Mohammadi-Shemirani, P., Shoamanesh, A., O'Donnell, M., Yusuf, S., Langhorne, P., & Par

1229 e, G. (2022). Mitochondrial DNA Copy Number as a Marker and Mediator of Stroke
1230 Prognosis. *Neurology*, *98*(5), e470–e482.
1231 <https://doi.org/10.1212/WNL.00000000000013165>

1232 Cochran, W. G. (1954). The Combination of Estimates from Different Experiments. *Biometrics*,
1233 *10*(1), 101. <https://doi.org/10.2307/3001666>

1234 Cui, R., Elzur, R. A., Kanai, M., Ulirsch, J. C., Weissbrod, O., Daly, M. J., Neale, B. M., Fan, Z., &
1235 Finucane, H. K. (2022). Improving fine-mapping by modeling infinitesimal effects. *BioRxiv*,
1236 2022.10.21.513123. <https://doi.org/10.1101/2022.10.21.513123>

1237 Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A.,
1238 Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and
1239 BCFtools. *GigaScience*, *10*(2), 1–4. <https://doi.org/10.1093/GIGASCIENCE/GIAB008>

1240 Dayama, G., Emery, S. B., Kidd, J. M., & Mills, R. E. (2014). The genomic landscape of
1241 polymorphic human nuclear mitochondrial insertions. *Nucleic Acids Research*, *42*(20),
1242 12640–12649. <https://doi.org/10.1093/nar/gku1038>

1243 de Bakker, P. I. W., Ferreira, M. A. R., Jia, X., Neale, B. M., Raychaudhuri, S., & Voight, B. F.
1244 (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association
1245 studies. *Human Molecular Genetics*, *17*(R2), R122–R128.
1246 <https://doi.org/10.1093/HMG/DDN288>

1247 Deming, W. E. (1943). Statistical adjustment of data. In *Statistical adjustment of data*. Wiley.

1248 D’Erchia, A. M., Atlante, A., Gadaleta, G., Pavesi, G., Chiara, M., de Virgilio, C., Manzari, C.,
1249 Mastropasqua, F., Prazzoli, G. M., Picardi, E., Gissi, C., Horner, D., Reyes, A., Sbisà, E., Tullo,
1250 A., & Pesole, G. (2015). Tissue-specific mtDNA abundance from exome data and its
1251 correlation with mitochondrial transcription, mass and respiratory activity. *Mitochondrion*,
1252 *20*, 13–21. <https://doi.org/10.1016/j.mito.2014.10.005>

1253 Ding, J., Sidore, C., Butler, T. J., Wing, M. K., Qian, Y., Meirelles, O., Busonero, F., Tsoi, L. C.,
1254 Maschio, A., Angius, A., Kang, H. M., Nagaraja, R., Cucca, F., Abecasis, G. çR, & Schlessinger,
1255 D. (2015). Assessing Mitochondrial DNA Variation and Copy Number in Lymphocytes of
1256 ~2,000 Sardinians Using Tailored Sequencing Analysis Tools. *PLoS Genetics*, *11*(7), 1005306.
1257 <https://doi.org/10.1371/JOURNAL.PGEN.1005306>

1258 Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high
1259 throughput. *Nucleic Acids Research*, *32*(5), 1792–1797.
1260 <https://doi.org/10.1093/NAR/GKH340>

1261 Ekstrand, M. I., Falkenberg, M., Rantanen, A., Park, C. B., Gaspari, M., Hultenby, K., Rustin, P.,
1262 Gustafsson, C. M., & Larsson, N. G. (2004). Mitochondrial transcription factor A regulates
1263 mtDNA copy number in mammals. *Human Molecular Genetics*, *13*(9), 935–944.
1264 <https://doi.org/10.1093/hmg/ddh109>

1265 Elliott, H. R., Samuels, D. C., Eden, J. A., Relton, C. L., & Chinnery, P. F. (2008). Pathogenic
1266 Mitochondrial DNA Mutations Are Common in the General Population. *American Journal*
1267 *of Human Genetics*, *83*(2), 254–260. <https://doi.org/10.1016/j.ajhg.2008.07.004>

1268 Falah, M., Farhadi, M., Kamrava, S. K., Mahmoudian, S., Daneshi, A., Balali, M., Asghari, A., &
1269 Houshmand, M. (2017). Association of genetic variations in the mitochondrial DNA control
1270 region with presbycusis. *Clinical Interventions in Aging*, *12*, 459.
1271 <https://doi.org/10.2147/CIA.S123278>

1272 Falkenberg, M., & Gustafsson, C. M. (2020). Mammalian mitochondrial DNA replication and
1273 mechanisms of deletion formation. <https://doi.org/10.1080/10409238.2020.1818684>,
1274 55(6), 509–524. <https://doi.org/10.1080/10409238.2020.1818684>

1275 Fazzini, F., Lamina, C., Raftopoulou, A., Koller, A., Fuchsberger, C., Pattaro, C., del Greco, F. M.,
1276 Döttelmayer, P., Fendt, L., Fritz, J., Meiselbach, H., Schönherr, S., Forer, L., Weissensteiner,
1277 H., Pramstaller, P. P., Eckardt, K. U., Hicks, A. A., & Kronenberg, F. (2021). Association of
1278 mitochondrial DNA copy number with metabolic syndrome and type 2 diabetes in 14 176
1279 individuals. *Journal of Internal Medicine*, 290(1), 190–202.
1280 <https://doi.org/10.1111/joim.13242>

1281 Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P. R., Anttila, V., Xu, H.,
1282 Zang, C., Farh, K., Ripke, S., Day, F. R., Purcell, S., Stahl, E., Lindstrom, S., Perry, J. R. B.,
1283 Okada, Y., Raychaudhuri, S., Daly, M. J., ... Price, A. L. (2015). Partitioning heritability by
1284 functional annotation using genome-wide association summary statistics. *Nature Genetics*,
1285 47(11), 1228–1235. <https://doi.org/10.1038/ng.3404>

1286 Finucane, H. K., Reshef, Y. A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.
1287 R., Lareau, C., Shores, N., Genovese, G., Saunders, A., Macosko, E., Pollack, S., Perry, J. R.
1288 B., Buenrostro, J. D., Bernstein, B. E., Raychaudhuri, S., McCarroll, S., ... Price, A. L. (2018).
1289 Heritability enrichment of specifically expressed genes identifies disease-relevant tissues
1290 and cell types. *Nature Genetics*, 50(4), 621–629. [https://doi.org/10.1038/s41588-018-](https://doi.org/10.1038/s41588-018-0081-4)
1291 0081-4

1292 Frazier, A. E., Thorburn, D. R., & Compton, A. G. (2019). Mitochondrial energy generation
1293 disorders: Genes, mechanisms, and clues to pathology. *Journal of Biological Chemistry*,
1294 294(14), 5386–5395. <https://doi.org/10.1074/jbc.R117.809194>

1295 Gazal, S., Finucane, H. K., Furlotte, N. A., Loh, P. R., Palamara, P. F., Liu, X., Schoech, A., Bulik-
1296 Sullivan, B., Neale, B. M., Gusev, A., & Price, A. L. (2017). Linkage disequilibrium-dependent
1297 architecture of human complex traits shows action of negative selection. *Nature Genetics*,
1298 49(10), 1421–1427. <https://doi.org/10.1038/ng.3954>

1299 Gitschlag, B. L., Kirby, C. S., Samuels, D. C., Gangula, R. D., Mallal, S. A., & Patel, M. R. (2016).
1300 Homeostatic Responses Regulate Selfish Mitochondrial Genome Dynamics in *C. elegans*.
1301 *Cell Metabolism*, 24(1), 91–103. <https://doi.org/10.1016/j.cmet.2016.06.008>

1302 Gopal, R. K., Calvo, S. E., Shih, A. R., Chaves, F. L., McGuone, D., Mick, E., Pierce, K. A., Li, Y.,
1303 Garofalo, A., van Allen, E. M., Clish, C. B., Oliva, E., & Mootha, V. K. (2018). Early loss of
1304 mitochondrial complex I and rewiring of glutathione metabolism in renal oncocytoma.
1305 *Proceedings of the National Academy of Sciences of the United States of America*, 115(27),
1306 E6283–E6290.
1307 https://doi.org/10.1073/PNAS.1711888115/SUPPL_FILE/PNAS.1711888115.SD05.XLSX

1308 Gopal, R. K., Kübler, K., Calvo, S. E., Polak, P., Livitz, D., Rosebrock, D., Sadow, P. M., Campbell,
1309 B., Donovan, S. E., Amin, S., Gigliotti, B. J., Grabarek, Z., Hess, J. M., Stewart, C., Braunstein,
1310 L. Z., Arndt, P. F., Mordecai, S., Shih, A. R., Chaves, F., ... McFadden, D. G. (2018).
1311 Widespread Chromosomal Losses and Mitochondrial DNA Alterations as Genetic Drivers in
1312 Hürthle Cell Carcinoma. *Cancer Cell*, 34(2), 242–255.e5.
1313 <https://doi.org/10.1016/J.CCELL.2018.06.013>

1314 Gupta, R., Karczewski, K. J., Howrigan, D., Neale, B. M., & Mootha, V. K. (2021). Human genetic
1315 analyses of organelles highlight the nucleus in age-related trait heritability. *ELife*, *10*,
1316 e68610. <https://doi.org/10.7554/eLife.68610>

1317 Hägg, S., Jylhävä, J., Wang, Y., Czene, K., & Grassmann, F. (2020). Deciphering the genetic and
1318 epidemiological landscape of mitochondrial DNA abundance. *Human Genetics*, *140*(6),
1319 849–861. <https://doi.org/10.1007/s00439-020-02249-w>

1320 Hail Team. (n.d.). *Hail 0.2*. Retrieved December 19, 2022, from <https://github.com/hail-is/hail>

1321 Halldorsson, B. v., Eggertsson, H. P., Moore, K. H. S., Hauswedell, H., Eiriksson, O., Ulfarsson, M.
1322 O., Palsson, G., Hardarson, M. T., Oddsson, A., Jensson, B. O., Kristmundsdottir, S.,
1323 Sigurpalsdottir, B. D., Stefansson, O. A., Beyter, D., Holley, G., Tragante, V., Gylfason, A.,
1324 Olason, P. I., Zink, F., ... Stefansson, K. (2022). The sequences of 150,119 genomes in the
1325 UK Biobank. *Nature* *2022* *607*:7920, 607(7920), 732–740. [https://doi.org/10.1038/s41586-](https://doi.org/10.1038/s41586-022-04965-x)
1326 [022-04965-x](https://doi.org/10.1038/s41586-022-04965-x)

1327 Holt, I. J., Harding, A. E., & Morgan-Hughes, J. A. (1988). Deletions of muscle mitochondrial DNA
1328 in patients with mitochondrial myopathies. *Nature*, *331*(6158), 717–719.
1329 <https://doi.org/10.1038/331717a0>

1330 Hormozdiari, F., van de Bunt, M., Segrè, A. v., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H.,
1331 Sankararaman, S., Pasaniuc, B., & Eskin, E. (2016). Colocalization of GWAS and eQTL Signals
1332 Detects Target Genes. *American Journal of Human Genetics*, *99*(6), 1245–1260.
1333 <https://doi.org/10.1016/j.ajhg.2016.10.003>

1334 Hurtado-Roca, Y., Ledesma, M., Gonzalez-Lazaro, M., Moreno-Loshuertos, R., Fernandez-Silva,
1335 P., Enriquez, J. A., & Laclaustra, M. (2016). Adjusting MtDNA quantification in whole blood
1336 for peripheral blood platelet and leukocyte counts. *PLoS ONE*, *11*(10).
1337 <https://doi.org/10.1371/journal.pone.0163770>

1338 Kanai, M., Elzur, R., Zhou, W., Daly, M. J., Finucane, H. K., Zhou, W., Kanai, M., Wu, K.-H. H.,
1339 Rasheed, H., Tsoo, K., Hirbo, J. B., Wang, Y., Bhattacharya, A., Zhao, H., Namba, S., Surakka,
1340 I., Wolford, B. N., lo Faro, V., Lopera-Maya, E. A., ... Neale, B. M. (2022). Meta-analysis fine-
1341 mapping is often miscalibrated at single-variant resolution. *Cell Genomics*, *2*(12), 100210.
1342 <https://doi.org/10.1016/J.XGEN.2022.100210>

1343 Kanai, M., Ulirsch, J. C., Karjalainen, J., Kurki, M., Karczewski, K. J., Fauman, E., Wang, Q. S.,
1344 Jacobs, H., Aguet, F., Ardlie, K. G., Kerimov, N., Alasoo, K., Benner, C., Ishigaki, K., Sakaue,
1345 S., Reilly, S., BioBank Japan Project, T., Kamatani, Y., Matsuda, K., ... Kanai mkanai, M.
1346 (2021). Insights from complex trait fine-mapping across diverse populations. *MedRxiv*,
1347 2021.09.03.21262975. <https://doi.org/10.1101/2021.09.03.21262975>

1348 Kang, E., Wu, J., Gutierrez, N. M., Koski, A., Tippner-Hedges, R., Agaronyan, K., Platero-Luengo,
1349 A., Martinez-Redondo, P., Ma, H., Lee, Y., Hayama, T., van Dyken, C., Wang, X., Luo, S.,
1350 Ahmed, R., Li, Y., Ji, D., Kayali, R., Cinnioglu, C., ... Mitalipov, S. (2016). Mitochondrial
1351 replacement in human oocytes carrying pathogenic mitochondrial DNA mutations. *Nature*,
1352 *540*(7632), 270–275. <https://doi.org/10.1038/nature20592>

1353 Karczewski, K. J., Solomonson, M., Chao, K. R., Goodrich, J. K., Tiao, G., Lu, W., Riley-Gillis, B. M.,
1354 Tsai, E. A., Kim, H. I., Zheng, X., Rahimov, F., Esmaeeli, S., Grundstad, A. J., Reppell, M.,
1355 Waring, J., Jacob, H., Sexton, D., Bronson, P. G., Chen, X., ... Neale, B. M. (2022). Systematic
1356 single-variant and gene-based association testing of thousands of phenotypes in

1357 394,841 UK Biobank exomes. *Cell Genomics*, 2(9), 100168.
1358 <https://doi.org/10.1016/J.XGEN.2022.100168>

1359 Kennedy, S. R., Salk, J. J., Schmitt, M. W., & Loeb, L. A. (2013). Ultra-Sensitive Sequencing
1360 Reveals an Age-Related Increase in Somatic Mitochondrial Mutations That Are Inconsistent
1361 with Oxidative Damage. *PLoS Genetics*, 9(9).
1362 <https://doi.org/10.1371/journal.pgen.1003794>

1363 Kerimov, N., Hayhurst, J. D., Peikova, K., Manning, J. R., Walter, P., Kolberg, L., Samoviča, M.,
1364 Sakthivel, M. P., Kuzmin, I., Trevanion, S. J., Burdett, T., Jupp, S., Parkinson, H.,
1365 Papatheodorou, I., Yates, A. D., Zerbino, D. R., & Alasoo, K. (2021). A compendium of
1366 uniformly processed human gene expression and splicing quantitative trait loci. *Nature*
1367 *Genetics* 2021 53:9, 53(9), 1290–1299. <https://doi.org/10.1038/s41588-021-00924-w>

1368 Laricchia, K. M., Lake, N. J., Watts, N. A., Shand, M., Haessly, A., Gauthier, L., Benjamin, D.,
1369 Banks, E., Soto, J., Garimella, K., Emery, J., Aggregation, G., Consortium, D., Rehm, H. L.,
1370 Macarthur, D. G., Tiao, G., Lek, M., Mootha, V. K., & Calvo, S. E. (2022). Mitochondrial DNA
1371 variation across 56,434 individuals in gnomAD. *Genome Res.* <https://bravo.sph.umich.edu>

1372 Li, M., Schröder, R., Ni, S., Madea, B., & Stoneking, M. (2015). Extensive tissue-related and
1373 allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations.
1374 *Proceedings of the National Academy of Sciences*, 112(8), 2491–2496.
1375 <https://doi.org/10.1073/pnas.1419651112>

1376 Li, M., Schroeder, R., Ko, A., & Stoneking, M. (2012). Fidelity of capture-enrichment for mtDNA
1377 genome sequencing: Influence of NUMTs. *Nucleic Acids Research*, 40(18).
1378 <https://doi.org/10.1093/nar/gks499>

1379 Lim, S. E., Longley, M. J., & Copeland, W. C. (1999). The mitochondrial p55 accessory subunit of
1380 human DNA polymerase γ enhances DNA binding, promotes processive DNA synthesis, and
1381 confers N-ethylmaleimide resistance. *Journal of Biological Chemistry*, 274(53), 38197–
1382 38203. <https://doi.org/10.1074/jbc.274.53.38197>

1383 Lin, Y. F., Schulz, A. M., Pellegrino, M. W., Lu, Y., Shaham, S., & Haynes, C. M. (2016).
1384 Maintenance and propagation of a deleterious mitochondrial genome by the
1385 mitochondrial unfolded protein response. *Nature* 2016 533:7603, 533(7603), 416–419.
1386 <https://doi.org/10.1038/nature17989>

1387 Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S., & Hirschhorn, J. N. (2003). Meta-analysis
1388 of genetic association studies supports a contribution of common variants to susceptibility
1389 to common disease. *Nature Genetics* 2003 33:2, 33(2), 177–182.
1390 <https://doi.org/10.1038/ng1071>

1391 Longchamps, R. J., Yang, S. Y., Castellani, C. A., Shi, W., Lane, J., Grove, M. L., Bartz, T. M.,
1392 Sarnowski, C., Burrows, K., Guyatt, A. L., Gaunt, T. R., Kacprowski, T., Yang, J., de Jager, P.
1393 L., Yu, L., Bergman, A., Xia, R., Fornage, M., Feitosa, M. F., ... Arking, D. E. (2021). Genome-
1394 wide analysis of mitochondrial DNA copy number reveals multiple loci implicated in
1395 nucleotide metabolism, platelet activation, and megakaryocyte proliferation. *BioRxiv*,
1396 2021.01.25.428086. <https://doi.org/10.1101/2021.01.25.428086>

1397 Longley, M. J., Clark, S., Man, C. Y. W., Hudson, G., Durham, S. E., Taylor, R. W., Nightingale, S.,
1398 Turnbull, D. M., Copeland, W. C., & Chinnery, P. F. (2006). Mutant POLG2 Disrupts DNA
1399 Polymerase γ Subunits and Causes Progressive External Ophthalmoplegia. *The American*
1400 *Journal of Human Genetics*, 78(6), 1026–1034. <https://doi.org/10.1086/504303>

1401 Lynch, M., Butcher, D., Bürger, R., & Gabriel, W. (1993). The Mutational Meltdown in Asexual
1402 Populations. *Journal of Heredity*, *84*(5), 339–344.
1403 <https://doi.org/10.1093/OXFORDJOURNALS.JHERED.A111354>

1404 Marchington, D. R., Hartshorne, G. M., Barlow, D., & Poulton, J. (1997). Homopolymeric tract
1405 heteroplasmy in mtDNA from tissues and single oocytes: support for a genetic bottleneck.
1406 *American Journal of Human Genetics*, *60*(2), 408.
1407 [/pmc/articles/PMC1712400/?report=abstract](https://pubmed.ncbi.nlm.nih.gov/1712400/)

1408 McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., &
1409 Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 1–
1410 14. <https://doi.org/10.1186/S13059-016-0974-4/TABLES/8>

1411 Nishino, I., Spinazzola, A., & Hirano, M. (1999). Thymidine Phosphorylase Gene Mutations in
1412 MNGIE, a Human Mitochondrial Disorder. *Science*, *283*(5402), 689–692.
1413 <https://doi.org/10.1126/science.283.5402.689>

1414 *Pan UKBB Initiative*. (2022). <https://pan.ukbb.broadinstitute.org/>

1415 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar,
1416 P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-
1417 Genome Association and Population-Based Linkage Analyses. *American Journal of Human*
1418 *Genetics*, *81*(3), 559. <https://doi.org/10.1086/519795>

1419 Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic
1420 features. *Bioinformatics*, *26*(6), 841–842. <https://doi.org/10.1093/bioinformatics/btq033>

1421 Rath, S., Sharma, R., Gupta, R., Ast, T., Chan, C., Durham, T. J., Goodman, R. P., Grabarek, Z.,
1422 Haas, M. E., Hung, W. H. W., Joshi, P. R., Jourdain, A. A., Kim, S. H., Kotrys, A. v, Lam, S. S.,
1423 Mccoy, J. G., Meisel, J. D., Miranda, M., Panda, A., ... Mootha, V. K. (2020). MitoCarta3.0:
1424 an updated mitochondrial proteome now with sub-organelle localization and pathway
1425 annotations. *Nucleic Acids Research*, 1–7. <https://doi.org/10.1093/nar/gkaa1011>

1426 Ratnaike, T. E., Greene, D., Wei, W., Sanchis-Juan, A., Schon, K. R., van den Aamee, J.,
1427 Raymond, L., Horvath, R., Turro, E., & Chinnery, P. F. (2021). MitoPhen database: A human
1428 phenotype ontology-based approach to identify mitochondrial DNA diseases. *Nucleic Acids*
1429 *Research*, *49*(17), 9686–9695. <https://doi.org/10.1093/nar/gkab726>

1430 Ryzdanicz, M., Cywińska, K., Wróbel, M., Pollak, A., Gawócki, W., Wojsyk-Banaszak, I.,
1431 Lechowicz, U., Mueller-Malesińska, M., Ołdak, M., Płoski, R., Skarzyński, H., Szyfter, K., &
1432 Szyfter, W. (2011). The contribution of the mitochondrial COI/tRNA Ser(UCN) gene
1433 mutations to non-syndromic and aminoglycoside-induced hearing loss in Polish patients.
1434 *Molecular Genetics and Metabolism*, *104*(1–2), 153–159.
1435 <https://doi.org/10.1016/j.ymgme.2011.05.004>

1436 Sang, Y., Liu, J. Y., Wang, F. Y., Luo, X. Y., Chen, Z. Q., Zhuang, S. M., & Zhu, Y. (2022).
1437 Mitochondrial micropeptide STMP1 promotes G1/S transition by enhancing mitochondrial
1438 complex IV activity. *Molecular Therapy*, *30*(8), 2844–2855.
1439 <https://doi.org/10.1016/J.YMTHE.2022.04.012>

1440 Sharma, R., Reinstadler, B., Engelstad, K., Skinner, O. S., Stackowitz, E., Haller, R. G., Clish, C. B.,
1441 Pierce, K., Walker, M. A., Fryer, R., Oglesbee, D., Mao, X., Shungu, D. C., Khatri, A., Hirano,
1442 M., de Vivo, D. C., & Mootha, V. K. (2021). Circulating markers of NADH-reductive stress
1443 correlate with mitochondrial disease severity. *Journal of Clinical Investigation*, *131*(2), 1–
1444 16. <https://doi.org/10.1172/JCI136055>

1445 Sherry, S. T., Ward, M., & Sirotkin, K. (1999). dbSNP—Database for Single Nucleotide
1446 Polymorphisms and Other Classes of Minor Genetic Variation. *Genome Research*, 9(8),
1447 677–679. <https://doi.org/10.1101/GR.9.8.677>

1448 Shoffner, J. M., Brown, M. D., Stugard, C., June, A. S., Pollock, S., Haas, R. H., Kaufman, A.,
1449 Koontz, D., Kim, Y., Graham, J. R., Smith, E., Dixon, J., & Wallace, D. C. (1995). Leber’s
1450 hereditary optic neuropathy plus dystonia is caused by a mitochondrial DNA point
1451 mutation. *Annals of Neurology*, 38(2), 163–169. <https://doi.org/10.1002/ANA.410380207>

1452 Stoneking, M. (2000). Hypervariable Sites in the mtDNA Control Region Are Mutational
1453 Hotspots. In *Am. J. Hum. Genet* (Vol. 67).

1454 Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green,
1455 J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T.,
1456 Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the
1457 Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, 12(3),
1458 1–10. <https://doi.org/10.1371/journal.pmed.1001779>

1459 Suomalainen, A., Kaukonen, J., Amati, P., Timonen, R., Haltia, M., Weissenbach, J., Zeviani, M.,
1460 Somer, H., & Peltonen, L. (1995). An autosomal locus predisposing to deletions of
1461 mitochondrial DNA. *Nature Genetics*, 9(2), 146–151. <https://doi.org/10.1038/ng0295-146>

1462 Tan, B. G., Wellesley, F. C., Savery, N. J., & Szczelkun, M. D. (2016). Length heterogeneity at
1463 conserved sequence block 2 in human mitochondrial DNA acts as a rheostat for RNA
1464 polymerase POLRMT activity. *Nucleic Acids Research*, 44(16), 7817–7829.
1465 <https://doi.org/10.1093/nar/gkw648>

1466 Terra. (n.d.). Retrieved November 4, 2022, from <https://app.terra.bio/>

1467 The “All of Us” Research Program. (2019). *New England Journal of Medicine*, 381(7), 668–676.
1468 https://doi.org/10.1056/NEJMSR1809937/SUPPL_FILE/NEJMSR1809937_APPENDIX.PDF

1469 Thomas, W. K., & Wilson, A. C. (1991). Mode and tempo of molecular evolution in the
1470 nematode *Caenorhabditis*: cytochrome oxidase II and calmodulin sequences. *Genetics*,
1471 128(2), 269–279. <https://doi.org/10.1093/GENETICS/128.2.269>

1472 Trifunovic, A., Wredenberg, A., Falkenberg, M., Spelbrink, J. N., Rovio, A. T., Bruder, C. E.,
1473 Bohlooly-Y, M., Gdlöf, S., Oldfors, A., Wibom, R., Törnell, J., Jacobs, H. T., & Larsson, N. G.
1474 (2004). Premature ageing in mice expressing defective mitochondrial DNA polymerase.
1475 *Nature*, 429(6990), 417–423. <https://doi.org/10.1038/nature02517>

1476 Uhler, J. P., Thörn, C., Nicholls, T. J., Matic, S., Milenkovic, D., Gustafsson, C. M., & Falkenberg,
1477 M. (2016). MGME1 processes flaps into ligatable nicks in concert with DNA polymerase γ
1478 during mtDNA replication. *Nucleic Acids Research*, 44(12), 5861–5871.
1479 <https://doi.org/10.1093/nar/gkw468>

1480 van Goethem, G., Dermaut, B., Löfgren, A., Martin, J.-J., & van Broeckhoven, C. (2001).
1481 Mutation of POLG is associated with progressive external ophthalmoplegia characterized
1482 by mtDNA deletions. *Nature Genetics*, 28(3), 211–212. <https://doi.org/10.1038/90034>

1483 Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). Springer.

1484 Walker, M. A., Lareau, C. A., Ludwig, L. S., Karaa, A., Sankaran, V. G., Regev, A., & Mootha, V. K.
1485 (2020). Purifying Selection against Pathogenic Mitochondrial DNA in Human T Cells. *New*
1486 *England Journal of Medicine*, 383(16), 1556–1563.
1487 <https://doi.org/10.1056/nejmoa2001265>

1488 Wallace, D. C., Singh, G., Lott, M. T., Hodge, J. A., Schurr, T. G., Lezza, A. M. S., Elsas, L. J., &
1489 Nikoskelainen, E. K. (1988). Mitochondrial DNA Mutation Associated with Leber's
1490 Hereditary Optic Neuropathy. *Science*, *242*(4884), 1427–1430.
1491 <https://doi.org/10.1126/SCIENCE.3201231>

1492 Wanagat, J., Cao, Z., Pathare, P., & Aiken, J. M. (2001). Mitochondrial DNA deletion mutations
1493 colocalize with segmental electron transport system abnormalities, muscle fiber atrophy,
1494 fiber splitting, and oxidative damage in sarcopenia. *FASEB Journal*, *15*(2), 322–332.
1495 <https://doi.org/10.1096/fj.00-0320com>

1496 Wang, G., Sarkar, A., Carbonetto, P., & Stephens, M. (2020). A simple new approach to variable
1497 selection in regression, with application to genetic fine mapping. *Journal of the Royal*
1498 *Statistical Society: Series B (Statistical Methodology)*, *82*(5), 1273–1300.
1499 <https://doi.org/10.1111/RSSB.12388>

1500 Wanrooij, P. H., Uhler, J. P., Shi, Y., Westerlund, F., Falkenberg, M., & Gustafsson, C. M. (2012).
1501 A hybrid G-quadruplex structure formed between RNA and DNA explains the extraordinary
1502 stability of the mitochondrial R-loop. *Nucleic Acids Research*, *40*(20), 10334.
1503 <https://doi.org/10.1093/NAR/GKS802>

1504 Wanrooij, P. H., Uhler, J. P., Simonsson, T., Falkenberg, M., & Gustafsson, C. M. (2010). G-
1505 quadruplex structures in RNA stimulate mitochondrial transcription termination and
1506 primer formation. *Proceedings of the National Academy of Sciences of the United States of*
1507 *America*, *107*(37), 16072–16077. <https://doi.org/10.1073/pnas.1006026107>

1508 Wei, W., Schon, K. R., Elgar, G., Orioli, A., Tanguy, M., Giess, A., Tischkowitz, M., Caulfield, M. J.,
1509 & Chinnery, P. F. (2022). Nuclear-embedded mitochondrial DNA sequences in 66,083
1510 human genomes. *Nature*. <https://doi.org/10.1038/s41586-022-05288-7>

1511 Weissensteiner, H., Forer, L., Fendt, L., Kheirikhah, A., Salas, A., Kronenberg, F., & Schoenherr, S.
1512 (2021). Contamination detection in sequencing studies using the mitochondrial phylogeny.
1513 *Genome Research*, *31*(2), 309–316. <https://doi.org/10.1101/GR.256545.119>

1514 Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H. J.,
1515 Kronenberg, F., Salas, A., & Schön herr, S. (2016). HaploGrep 2: mitochondrial haplogroup
1516 classification in the era of high-throughput sequencing. *Nucleic Acids Research*, *44*(W1),
1517 W58–W63. <https://doi.org/10.1093/nar/gkw233>

1518 Xuan, H. P., Farge, G., Shi, Y., Gaspari, M., Gustafsson, C. M., & Falkenberg, M. (2006).
1519 Conserved sequence box II directs transcription termination and primer formation in
1520 mitochondria. *Journal of Biological Chemistry*, *281*(34), 24647–24652.
1521 <https://doi.org/10.1074/jbc.M602429200>

1522 Yang, S. Y., Castellani, C. A., Longchamps, R. J., Pillalamarri, V. K., O'Rourke, B., Guallar, E., &
1523 Arking, D. E. (2021). Blood-derived mitochondrial DNA copy number is associated with
1524 gene expression across multiple tissues and is predictive for incident neurodegenerative
1525 disease. *Genome Research*, *31*(3), 349–358. <https://doi.org/10.1101/GR.269381.120>

1526 Zhou, W., Kanai, M., Wu, K. H. H., Rasheed, H., Tsoo, K., Hirbo, J. B., Wang, Y., Bhattacharya, A.,
1527 Zhao, H., Namba, S., Surakka, I., Wolford, B. N., lo Faro, V., Lopera-Maya, E. A., Läll, K.,
1528 Favé, M. J., Partanen, J. J., Chapman, S. B., Karjalainen, J., ... Neale, B. M. (2022). Global
1529 Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell*
1530 *Genomics*, *2*(10), 100192. <https://doi.org/10.1016/J.XGEN.2022.100192>

1531 Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wofford, B. N., LeFaive, J.,
1532 VandeHaar, P., Gagliano, S. A., Gifford, A., Bastarache, L. A., Wei, W. Q., Denny, J. C., Lin,
1533 M., Hveem, K., Kang, H. M., Abecasis, G. R., Willer, C. J., & Lee, S. (2018). Efficiently
1534 controlling for case-control imbalance and sample relatedness in large-scale genetic
1535 association studies. *Nature Genetics* 2018 50:9, 50(9), 1335–1341.
1536 <https://doi.org/10.1038/s41588-018-0184-y>
1537
1538