Genetic analysis of mitochondrial DNA copy number and associated traits identifies loci implicated in nucleotide metabolism, platelet activation, and megakaryocyte proliferation, and reveals a causal association of mitochondrial function with mortality

Longchamps RJ[1*], Yang SY[1*], Castellani CA[1,2], Shi W[1], Lane J[3], Grove ML[4], Bartz TM[5], Sarnowski C[6], Burrows K[7,8], Guyatt AL[9], Gaunt TR[7,8], Kacprowski T[10], Yang J[11], De Jager PL[12,13], Yu L[11], CHARGE Aging and Longevity Group, Bergman A[14], Xia R[15], Fornage M[15,16], Feitosa MF[17], Wojczynski MK[17], Kraja AT[17], Province MA[17], Amin N[18], Rivadeneira F[19], Tiemeier H[18,20], Uitterlinden AG[18,19], Broer L[19], Van Meurs JBJ[18,19], Van Duijn CM[18], Raffield LM[21], Lange L[22], Rich SS[23], Lemaitre RN[24], Goodarzi MO[25], Sitlani CM[24], Mak ACY[26], Bennett DA[11], Rodriguez S[7,8], Murabito JM[27], Lunetta KL[6], Sotoodehnia N[28], Atzmon G[29], Kenny Y[30], Barzilai N[31], Brody JA[32], Psaty BM[33], Taylor KD[34], Rotter JI[34], Boerwinkle E[4,35], Pankratz N[3], Arking DE[1]

*** Indicates authors contributed equally to this work.**

1 McKusick-Nathans Institute, Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD

2 Department of Pathology and Laboratory Medicine, Western University, London, ON, Canada

3 Department of Laboratory Medicine and Pathology, University of Minnesota Medical School, Minneapolis, MN

4 Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX

5 Cardiovascular Health Research Unit, Departments of Medicine and Biostatistics, University of Washington, Seattle, WA

6 Department of Biostatistics, Boston University School of Public Health, Boston, MA

24 7 MRC Integrative Epidemiology Unit at the University of Bristol, University of Bristol, Oakfield House,
25 Oakfield Grove, Bristol, UK
26 8 Population Health Sciences, Bristol Medical School, University of Bristol, Oakfield House, Oakfield
27 Grove, Bristol, UK
28 9 Department of Health Sciences, University of Leicester, University Road, Leicester, UK
29 10 Department of Functional Genomics, Interfaculty Institute for Genetics and Functional Genomics,
30 University of Greifswald, Greifswald, Germany; Division Data Science in Biomedicine, Peter L. Reichertz
31 Institute for Medical Informatics, TU Braunschweig and Hannover Medical School, Brunswick, Germany
32 11 Rush Alzheimer's Disease Center & Department of Neurological Sciences, Rush University Medical
33 Center, Chicago, IL
34 12 Center for Translational and Systems Neuroimmunology, Department of Neurology, Columbia
35 University Medical Center, New York, NY
36 13 Program in Medical and Population Genetics, Broad Institute, Cambridge, MA
37 14 Department of Systems and Computational Biology, Albert Einstein College of Medicine, Bronx, New
38 York, USA
39 15 Institute of Molecular Medicine, The University of Texas health Science Center at Houston, Houston
40 TX
41 16 Human Genetics Center, The University of Texas Health Science Center at Houston, Houston
42 17 Division of Statistical Genomics, Department of Genetics, Washington University School of Medicine
43 18 Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands
44 19 Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands
45 20 Department of Social and Behavioral Science, Harvard T.H. School of Public Health, Boston, USA
46 21 Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC
47 22 Department of Medicine, University of Colorado Denver, Anschutz Medical Campus, Aurora, CO

48 23 Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA
49 24 Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle,
50 WA
51 25 Division of Endocrinology, Diabetes and Metabolism, Cedars-Sinai Medical Center, Los Angeles, CA
52 26 Cardiovascular Research Institute and Institute for Human Genetics, University of California, San
53 Francisco, California
54 27 Boston University School of Medicine, Boston University, Boston, MA
55 28 Cardiovascular Health Research Unit, Division of Cardiology, University of Washington, Seattle, WA
56 29 Department of Natural science, University of Haifa, Haifa, Israel; Departments of Medicine and
57 Genetics, Albert Einstein College of Medicine, Bronx, NY, 10461, USA.
58 30 Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY,
59 10461, USA.
60 31 Departments of Medicine and Genetics, Albert Einstein College of Medicine, Bronx, NY, 10461, USA.
61 32 Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle,
62 WA
63 33 Cardiovascular Health Research Unit, Departments of Epidemiology, Medicine and Health Services,
64 University of Washington, Seattle, WA
65 34 The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The
66 Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA
67 35 Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX
68

69 **Abstract**

70 Mitochondrial DNA copy number (mtDNA-CN) measured from blood specimens is a minimally
71 invasive marker of mitochondrial function that exhibits both inter-individual and intercellular variation.

To identify genes involved in regulating mitochondrial function, we performed a genome-wide association study (GWAS) in 465,809 White individuals from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium and the UK Biobank (UKB). We identified 133 SNPs with statistically significant, independent effects associated with mtDNA-CN across 100 loci. A combination of fine-mapping, variant annotation, and co-localization analyses were used to prioritize genes within each of the 133 independent sites. Putative causal genes were enriched for known mitochondrial DNA depletion syndromes ($p = 3.09 \times 10^{-15}$) and the gene ontology (GO) terms for mtDNA metabolism ($p = 1.43 \times 10^{-8}$) and mtDNA replication ($p = 1.2 \times 10^{-7}$). A clustering approach leveraged pleiotropy between mtDNA-CN associated SNPs and 41 mtDNA-CN associated phenotypes to identify functional domains, revealing three distinct groups, including platelet activation, megakaryocyte proliferation, and mtDNA metabolism. Finally, using mitochondrial SNPs, we establish causal relationships between mitochondrial function and a variety of blood cell related traits, kidney function, liver function and overall ($p = 0.044$) and non-cancer mortality ($p = 6.56 \times 10^{-4}$).

Introduction

Mitochondria are the cellular organelles primarily responsible for producing the chemical energy required for metabolism, as well as signaling the apoptotic process, maintaining homeostasis, and synthesizing several macromolecules such as lipids, heme and iron-sulfur clusters[1,2]. Mitochondria possess their own genome (mtDNA); a circular, intron-free, double-stranded, haploid, ~16.6 kb maternally inherited molecule encoding 37 genes vital for proper mitochondrial function. Due to the integral role of mitochondria in cellular metabolism, mitochondrial dysfunction is known to play a critical role in the underlying etiology of several aging-related diseases[3–5].

95 Unlike the nuclear genome, a large amount of variation exists in the number of copies of mtDNA present
96 within cells, tissues, and individuals. The relative copy number of mtDNA (mtDNA-CN) has been shown
97 to be positively correlated with oxidative stress[6], energy reserves, and mitochondrial membrane
98 potential[7]. As a minimally invasive proxy measure of mitochondrial dysfunction[8], decreased mtDNA-CN
99 measured in blood has been previously associated with aging-related disease states including frailty[9],
100 cardiovascular disease[10–12], chronic kidney disease[13], neurodegeneration[14,15], and cancer[16].
101
102 Although mtDNA-CN measured from whole blood presents itself as an easily accessible and minimally
103 invasive biomarker, cell type composition has been shown to be an important confounder, complicating
104 analyses[17,18]. For example, while platelets generally have fewer mtDNA molecules than leukocytes, the
105 lack of a platelet nuclear genome drastically skews mtDNA-CN estimates. As a result, not only is
106 controlling for cell composition extremely vital for accurate mtDNA-CN estimation, interpreting the
107 results in relation to the impact of cell composition becomes a necessity[18–20].
108
109 Although the comprehensive mechanism through which mtDNA-CN is modulated is largely unknown[21,22],
110 twin studies have estimated a broad-sense heritability of ~0.65, consistent with moderate genetic
111 control[23]. Several nuclear genes have been shown to directly modulate mtDNA-CN, specifically those
112 within the mtDNA replication machinery such as the mitochondrial polymerase, *POLG* and *POLG2*[24,25], as
113 well as the mitochondrial DNA helicase, *TWNK*, and the mitochondrial single-stranded binding protein,
114 *mtSSB*[26]. Furthermore, nuclear genes which maintain proper mitochondrial nucleotide supply including
115 *DGUOK* and *TK2* have also been shown to regulate mtDNA-CN[27–29]. To further elucidate the genetic
116 control over mtDNA-CN, several genome-wide association studies (GWAS) of mtDNA-CN have been
117 published[30–33], including a study that was published while the current manuscript was in preparation,
118 analyzing ~300,000 participants from the UK Biobank (UKB), and identifying 50 independent loci[33].

119

120 In the present study, we report mtDNA-CN GWAS results from 465,809 individuals across the Cohorts
121 for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium[34] and the UK Biobank
122 (UKB)[35]. Using multiple gene prioritization and functional annotation methods, we assign genes to loci
123 that reach genome-wide significance. We perform a PHEWAS and group our genome-wide significant
124 SNPs into 3 clusters that represent distinct functional domains related to mtDNA-CN. Finally, we
125 leverage mitochondrial SNPs to establish causality between mitochondrial function and mtDNA-CN
126 associated traits.

127

128 **Subjects and Methods**

129

130 **Study Populations**

131 470,579 individuals participated in this GWAS, 465,809 of whom self-identified as White. Participants
132 were derived from 7 population-based cohorts representing the Cohorts for Heart and Aging Research in
133 Genetic Epidemiology (CHARGE) consortium (Avon Longitudinal Study of Parents and Children [ALSPAC],
134 Atherosclerosis Risk in Communities [ARIC], Cardiovascular Health Study [CHS], Multi-Ethnic Study of
135 Atherosclerosis [MESA], Religious Orders Study and Memory and Aging Project [ROSMAP], Study of
136 Health in Pomerania [SHIP]) and from the UK Biobank (UKB) (Supplemental Table 1). Detailed
137 descriptions of each participating cohort, their quality control practices, study level analyses, and ethic
138 statements are available in the Supplemental Methods. All study participants provided written informed
139 consent and all centers obtained approval from their institutional review boards.

140

141 **Methods for Mitochondrial DNA Copy Number Estimation (CHARGE cohorts)**

142 ***qPCR***

143 mtDNA-CN was determined using a quantitative PCR assay as previously described[32,36]. Briefly, the cycle
144 threshold (Ct) value of a nuclear-specific and mitochondrial-specific probe were measured in triplicate
145 for each sample. In CHS, a multiplex assay using the mitochondrial *ND1* probe and nuclear *RPPH1* probe
146 was used, whereas ALSPAC used a mitochondrial probe targeting the D-Loop and a nuclear probe
147 targeting *B2M*. In CHS, we observed plate effects, as well as a linear increase in ΔCt due to the pipetting
148 order of each replicate. These effects were corrected in the analysis using linear mixed model
149 regression, with pipetting order included as a fixed effect and plate as a random effect to create a raw
150 measure of mtDNA-CN. In ALSPAC, run-to-run variability was controlled using 3 calibrator samples
151 added to every plate, to allow for adjustment by a per-plate calibration factor[32].
152
153 ***Microarray***
154 Microarray probe intensities were used to estimate mtDNA-CN using the Genvisis software package[37] as
155 previously described[10,36]. Briefly, Genvisis uses the median mitochondrial probe intensity across all
156 homozygous mitochondrial SNPs as an initial estimate of mtDNA-CN. Technical artifacts such as DNA
157 input quality, DNA input quantity, and hybridization efficiency were captured through either surrogate
158 variable (SV) or principal component (PC) analyses. SVs or PCs were adjusted for through stepwise linear
159 regression by adding successive components until each successive surrogate variable or principal
160 component no longer significantly improved the model.
161
162 ***Whole Genome Sequencing (ARIC)***
163 Whole genome sequencing read counts were used to estimate mtDNA-CN as previously described[36].
164 Briefly, the total number of reads in a sample were web scraped from the NCBI sequence read archive.
165 Mitochondrial reads were downloaded directly from dbGaP through Samtools (1.3.1). There was no

166 overlap between ARIC microarray and ARIC whole-genome sequencing samples. A ratio of mitochondrial
167 reads to total aligned reads was used as a raw measure of mtDNA-CN.
168
169 ***Adjusting for Covariates***
170 Each method described above represents a raw measure of mtDNA-CN, adjusted for technical artifacts;
171 however, several potential confounding variables (e.g., age, sex, blood cell composition) have been
172 identified previously[18]. Raw mtDNA-CN values were adjusted for white blood cell count in ARIC, SHIP and
173 CHS (which also adjusted for platelet count), depending on available data. Standardized residuals (mean
174 = 0, standard deviation = 1) of mtDNA-CN were used after adjusting for covariates (Supplemental Table
175 1).
176
177 **Estimation of Mitochondrial DNA Copy Number (UKB)**
178 Due to the availability of more detailed cell count data, as well as a different underlying biochemistry for
179 the Affymetrix Axiom array compared to the genotyping arrays used in the CHARGE cohorts, mtDNA-CN
180 in the UKB was estimated differently (Supplemental Methods). Briefly, mtDNA-CN estimates derived
181 from whole exome sequencing data, available on ~50,000 individuals, were generated first using
182 customized Perl scripts to aggregate the number of mapped sequencing reads and correct for covariates
183 through both linear and spline regression models. Concurrently, mitochondrial probe intensities from
184 the Affymetrix Axiom arrays, available on the full ~500,000 UKB cohort, were adjusted for technical
185 artifacts through principal components generated from nuclear probe intensities. Probe intensities were
186 then regressed onto the whole exome sequencing mtDNA-CN metric, and beta estimates from that
187 regression were used to estimate mtDNA-CN in the full UKB cohort. Finally, we used a 10-fold cross
188 validation method to select the cell counts to include in the final model (Supplemental Table 2). The

189 final UKB mtDNA-CN metric is the standardized residuals (mean = 0, standard deviation = 1) from a
190 linear model adjusting for covariates (age, sex, cell counts) as described in the Supplemental Methods.
191
192 **Genome-Wide Association Study**
193 For each individual cohort, regression analysis was performed with residualized mtDNA-CN as the
194 dependent variable adjusting for age, sex, and cohort-specific covariates (e.g., principal components,
195 DNA collection site, family structure, cell composition). Cohorts with multiple mtDNA-CN estimation
196 platforms were stratified into separate analyses. Ancestry-stratified meta-analyses were performed
197 using Metasoft software using the Han and Eskin random effects model to control for unobserved
198 heterogeneity due to differences in mtDNA-CN estimation method[38]. Effect size estimates for SNPs
199 were calculated using a random effect meta-analysis from cohort summary statistics, as the Han and
200 Eskin model relaxes the assumption under the null hypothesis without modifying the effect size
201 estimates that occur under the alternative hypothesis[38]. In total, three complementary analyses were
202 performed in self-identified White individuals: (1) a meta-analysis using all available studies, (2) a meta-
203 analysis of studies with available data for cell count adjustments, and (3) an analysis of UKB-only data.
204 As the vast majority of samples are derived from the UKB study, and given the difficulty in interpreting
205 effect size estimates from a random effects model, further downstream analyses were all performed
206 using effect size estimates from UKB-only data. We additionally performed X chromosome analyses,
207 using only UKB data. X-chromosome analyses were stratified by sex (males = 194,151, females =
208 216,989), and summary statistics were meta-analyzed using METAL[39] to obtain the final effect estimates.
209
210 **SNP Heritability Estimation**
211 SNP heritability estimates were retrieved from BOLT-LMM[40]. To verify this metric, we used SumHer[41] to
212 calculate an independent heritability metric using summary statistics. The heritability model used in this

213 analysis was the BLD-LDAK model. The tagging file used is the pre-computed UK Biobank GBR version for
214 the corresponding heritability model. The summary statistics were filtered so that only single-character
215 reference and alternate alleles are allowed. Chr:BP combination duplicates were removed except the
216 first appearance. SNP heritability was then calculated and extracted from output files.
217
218 **Identification of Independent GWAS Loci**
219 To identify the initial genome-wide significant (lead) SNPs in each locus, the most significant SNP that
220 passed genome-wide significance ($p < 5 \times 10^{-8}$) within a 1 Mb window was selected. To avoid Type I
221 error, SNPs were only retained for further analyses if there were either (a) at least two genome-wide
222 significant SNPs in the 1 Mb window or (b) if the lead SNP was directly genotyped. Conditional analyses
223 were performed in UKB, where the lead SNPs from the original GWAS were used as additional covariates
224 in order to identify additional independent associations.
225
226 **Comparisons with Hägg et al. 2020**
227 To compare results with Hägg et al. 2020[33], summary statistics were obtained from their Supplementary
228 Table 4. Loci were identified as shared between the two GWAS if two lead SNPs were fewer than
229 500,000 base pairs apart from one another.
230
231 **Fine-mapping**
232 The susieR package was used to identify all potential causal variants for each independent locus
233 associated with mtDNA CN[42]. UKB imputed genotype data for unrelated White subjects were used and
234 variants were extracted using a 500 kb window around the lead SNP for each locus with minor allele
235 frequency (MAF) > 0.001. 95% credible sets (CS) of SNPs, containing a potential causal variant within a
236 locus, were generated. The minimum absolute correlation within each CS is 0.5 and the scaled prior

237 variance is 0.01. When the CS did not include the lead SNP identified from the GWAS, some of the
238 parameters were slightly relaxed [minimum absolute correlation is 0.2, estimate prior variance is TRUE].
239 The SNP with the highest posterior inclusion probability (PIP) within each CS was also identified
240 (Supplemental Table 3). With a few exceptions, final lead SNPs were selected by prioritizing initially
241 identified SNPs unless the SNP with the highest PIP had a PIP greater than 0.2 and was 1.75 times larger
242 than the SNP with the second highest PIP.
243
244 **Functional Annotation and Gene Prioritization**
245 ***Functional Annotation***
246 ANNOVAR was used for functional annotation of variants identified in the fine-mapping step[43]. First,
247 variants were converted to an ANNOVAR-ready format using the dbSNP version 150 database[44]. Then,
248 variants were annotated with ANNOVAR using the RefSeq Gene database[45]. The annotation for each
249 variant includes the associated gene and region (e.g., exonic, intronic, intergenic). For intergenic
250 variants, ANNOVAR provides flanking genes and the distance to each gene. For exonic variants,
251 annotations also include likely functional consequences (e.g., synonymous/nonsynonymous,
252 insertion/deletion), the gene affected by the variant, and the amino acid sequence change
253 (Supplemental Table 4).
254 ***Co-localization Analyses***
255 Co-localization analyses were performed using the approximate Bayes factor method in the R package
256 *coloc*[46]. Briefly, *coloc* utilizes eQTL data and GWAS summary statistics to evaluate the probability that
257 gene expression and GWAS data share a single causal SNP (colocalize). *Coloc* returns multiple posterior
258 probabilities; H0 (no causal variant), H1 (causal variant for gene expression only), H2 (causal variant for
259 mtDNA-CN only), H3 (two distinct causal variants), and H4 (shared causal variant for gene expression
260 and mtDNA-CN). In the event of high H4, we designate the gene as causal for the GWAS phenotype of

261 interest (mtDNA-CN). eQTL summary statistics were obtained from the eQTLGen database[47]. Genes with
262 significant associations with lead SNPs were tested for co-localization using variants within a 500 kb
263 window of the sentinel SNP. Occasionally, some of the eQTLGen p-values for certain SNPs were
264 identical due to R's (ver 4.0.3) limitation in handling small numbers. To account for this, if the absolute
265 value for a SNP's z-score association with a gene was greater than 37.02, z-scores were rescaled so that
266 the largest z-score would result in a p-value of 5×10^{-300}. Additionally, a few clearly co-localized genes
267 did not result in high H4 PPs due to the strong effect for each phenotype of a single SNP (Supplemental
268 Figure 1), possibly due to differences in linkage disequilibrium (LD) between the UKB and eQTLGen
269 populations. To account for this, we summed mtDNA-CN GWAS p-values and eQTLGen p-values for each
270 SNP and removed the SNP with the lowest combined p-value. Co-localization analyses were then
271 repeated without the lowest SNP. Genes with H4 greater than 50% were classified as genes with
272 significant evidence of co-localization.

273 ***DEPICT***

274 Gene prioritization was performed with Depict, an integrative tool that incorporates gene co-regulation
275 and GWAS data to identify the most likely causal gene at a given locus[48]. Across GWAS SNPs which
276 overlapped with the DEPICT database, we identified SNPs representing 119 independent loci with LD
277 pruning defined as $p < 5 \times 10^{-8}$, $r^2 < 0.05$ and > 500 kb from other locus boundaries. Only genes with a
278 nominal p-value of less than 0.05 were considered for downstream prioritization.

279 ***Gene Assignment***

280 To prioritize genes for each identified locus, we utilized functional annotations, eQTL co-localization
281 analyses, and DEPICT gene prioritization results (Supplemental Figure 2). First, genes with missense
282 variants within susieR fine-mapped credible sets were assigned to loci. If loci co-localized with a gene's
283 expression with a posterior probability (PP) of greater than 0.50 and there were no other co-localized
284 genes with a PP within 5%, the gene with the highest posterior probability was assigned. If there was still

285 no assigned gene, the most significant DEPICT gene was assigned. If there was no co-localization or
286 DEPICT evidence, the nearest gene was assigned.
287
288 **Gene Set Enrichment Analyses**
289 Using the finalized gene list from the prioritization pipeline, GO and KEGG pathway enrichment analyses
290 were performed using the "goana" and "kegga" functions from the R package *limma*[49], treating all
291 known genes as the background universe[50]. Only one gene per locus was used for "goana" and "kegga"
292 gene set enrichment analysis, prioritizing genes assigned to primary independent hits. If there were
293 multiple assigned genes, one gene was randomly selected to avoid biasing results through loci with
294 multiple functionally related genes. To identify an appropriate p-value cutoff, 100 genes were randomly
295 selected from the genome and run through the same enrichment analysis. This permutation was
296 repeated 1000 times to generate a null distribution of the smallest p-values from each permutation. For
297 cluster-specific gene set enrichment analyses, permutation testing used the same number of random
298 genes as the number of genes in each cluster. To ensure robustness of results, gene set enrichment
299 analysis was repeated 50 times with random selection of genes at loci with multiple assigned genes. GO
300 and KEGG terms that passed permutation cutoffs at least 40/50 times were retained.
301
302 **Gene-based Association Test**
303 We used metaXcan, which employs gene expression prediction models to evaluate associations
304 between phenotypes and gene expression[51]. We obtained pre-calculated expression prediction models
305 and SNP covariance matrices, computed using whole blood from European ancestry individuals in
306 version 7 of the Genotype-Tissue expression (GTEx) database[52]. Using prediction performance $p < 0.05$, a
307 total of 6,285 genes were predicted. Of these genes, 74 passed Bonferroni correction of $p < 7.95 \times 10^{-6}$.
308 Gene set enrichment analyses were performed on Bonferroni-significant genes as previously described.

309 REVIGO[53] was used on the "medium" setting (allowed similarity = 0.7) to visualize significantly enriched
310 GO terms.
311
312 We used a one-sided Fisher's exact test to test for enrichment of genes that have been previously
313 identified as causal for mtDNA depletion syndromes[54–56].
314
315 **PHEWAS-based SNP Clustering**
316 ***mtDNA-CN Phenome-wide Association Study (PHEWAS)***
317 We used the PHEnome Scan ANalysis Tool (PHESANT)[57] to identify mtDNA-CN associated quantitative
318 traits in the UKB. Briefly, we tested for the association of mtDNA-CN with 869 quantitative traits
319 (Supplemental Table 5), limiting analyses to 365,781 White, unrelated individuals
320 (used.in.pca.calculation=1). As extreme cell count measurements could indicate individuals with active
321 infections or cancers, they were excluded from analysis (see Supplemental Methods). Analyses were
322 adjusted for age, sex, and assessment center.
323
324 ***SNP-Phenotype Associations***
325 SNP genotypes were regressed on mtDNA-associated quantitative phenotypic traits using linear
326 regression, adjusted for sex, age with a natural spline (df = 2), assessment center, genotyping array, and
327 40 genotyping principal components (provided as part of the UKB data download).
328
329 ***SNP Clustering***
330 To identify distinct clusters of mtDNA-CN GWS SNPs based on phenotypic associations, beta estimates
331 from the SNP-phenotype associations were first divided by the beta estimate of the mtDNA-CN SNP-
332 mtDNA-CN association, so that all SNP-phenotype associations are relative to the mtDNA-CN increasing

333 allele and scaled to the effect of the SNP on mtDNA-CN. The adjusted beta estimates were subjected to
334 a dimensionality reduction method, Uniform Manifold and Approximation Projection (UMAP), as
335 implemented in the R package *umap*[58] (random_state=123, n_neighbors=10, min_dist=0.001,
336 n_components=2, n_epochs=200). SNPs were assigned to clusters using Density Based Clustering of
337 Applications with Noise (DBSCAN) as implemented in the R package *dbscan*[59] (minPts=10). Robustness of
338 cluster assignment was established by varying n_neighors, min_dist, and random_state parameters.
339 Clusters represent groups of SNPs with similar phenotypic associations.
340

341 ***Phenotype Enrichment and Permutation Testing***

342 To test for enrichment of *individual* phenotypes within clusters, we compared the median mtDNA-CN
343 scaled phenotype beta estimates within the cluster to the median beta estimates for all SNPs not in the
344 cluster, with significance determined using 20,000 permutations in which cluster assignment was
345 permuted. For multi-test correction *across all* phenotypes, we performed 300 permutations of the initial
346 cluster assignment, followed by the comparison of median beta estimates as described above. We
347 retained only the most significant result from across all phenotypes and clusters from each of the 300
348 permutations, and then selected the 15th most significant value as the study-wide threshold for multi-
349 test corrected significance of $p < 0.05$.
350

351 **mtDNA Variant Association Analyses**

352 ***Mitochondrial Variant Phasing and Imputation***

353 Shapeit4 and Impute5 were used for UK Biobank mtDNA genotype phasing and imputation[60,61]. Phasing
354 and imputation were performed separately for each genotyping array (UKBB, UKBL), and restricted to
355 self-identified White individuals. The reference panel used for imputation analysis was the 1000
356 Genomes Project phase 3 mtDNA variants[62]. UK Biobank genotypes were coded to match the reference

357 panel allele. All genotype files, including the reference panel, were phased using Shapeit4 to fill in any
358 missing genotypes using the phasing iteration sequence "10b,1p,1b,1p,1b,1p,1b,1p,10m", where b is
359 burn-in iteration, p is pruning iteration, and m is main iteration. The –sequencing option was also used
360 due to the presence of multiple mtDNA variants in a very small region, analogous to sequencing data.
361 Phased UK Biobank genotypes were then imputed with the reference panel using Impute5 with
362 the following parameters: –pbwt-depth 8; –pbwt-cm 0.005; –no-threshold. All imputed variants were
363 functionally annotated using MSeqDR mvTools[63].
364
365 **mtSNP Association Tests**
366 Linear regressions stratified by genotyping array (UKBB, UKBL) were performed for each mtDNA SNP on
367 the 41 traits and mtDNA-CN, including the following covariates: age, age^2, sex, center, first 20
368 genotyping PCs. Only SNPs with MAF > 0.005 and imputation INFO score > 0.80 were included (UKBB,
369 n=223; UKBL, n=190; both, n=149). Results were then meta-analyzed using inverse variance weighting.
370 For association analyses between mitochondrial SNPs/haplotypes and mtDNA-CN, the mtDNA-CN metric
371 used was derived only from WES data, as mtDNA genotypes can subtly influence the estimates from
372 genotype array intensities.
373
374 **Identification of Independent Genetic Effects**
375 Single SNP study-wide significance was established by generating 300 normally distributed dummy
376 traits, and running single SNP tests using the UKBB data. The minimum SNP p-value for each dummy
377 trait was then selected, and the 15th most significant p-value from the 300 analyses was divided by 42
378 (41 real traits + mtDNA-CN), resulting in study-wide p-value threshold of $p < 9.5\text{x}10^{-6}$. To identify a
379 subset of traits to perform credible set identification using SusieR (see above, **Fine Mapping**), SNPs were
380 first filtered based on the study-wide p-value threshold, and then then most significantly associated trait

381 was identified for each SNP. SusieR, (parameters: L=10, estimate_residual_variance=TRUE,
382 estimate_prior_variance=TRUE, check_z = FALSE) was then run for each of these traits using the UKBB
383 imputed data and summary association test statistics. A total of 7 credible sets were identified across
384 the 4 traits, two of which co-localized, resulting in 6 credible sets. Independence across the 6 credible
385 sets was tested using multivariate regression models, and requiring $p < 0.0005$ for at least one trait for a
386 SNP to remain in the model. SNPs MT73A_G and MT 7028C_T were in moderate high LD (r^2=0.67), but
387 based on conditional regression analyses as described in the main results, capture independent effects
388 and are associated with different traits.
389
390 **Haplotype Generation and Analysis**
391 Haplotype were constructed by concatenating SNPs across the 6 credible sets using SNPs directly
392 genotyped on both genotyping arrays. This required selecting a SNP with a lower PIP for 2 of the 6
393 credible sets (MT12612A_G replaced MT462C_T, r^2 = 0.81; MT10238T_C replaced MT4529A_T, r^2 =
394 0.89). Haplotypes with MAF < 0.005 were set to missing (n = 1607), resulting in 8 haplotypes, with the
395 most common haplotype set as reference. Significance for haplotype associations with each trait were
396 generated by an anova between regression models with and without the haplotypes. Covariates
397 included age, age^2, sex, center, first 40 genotyping PCs, and genotyping array.
398 Mortality analyses were run using Cox proportional hazards models, with covariates as above.
399 Individuals with external causes of death (ICD 10 Death Code categories V, W, X, Y) were censored at
400 time of death. Additionally, for non-cancer mortality analyses, cancer death (ICD 10 Death Code
401 categories C00-D48) were censored. For cancer mortality analyses, all death due to non-cancer cases
402 were censored at time of death.
403

404 Clustering for visualization was performed using the R package 'heatmaply', with default setting and
405 hclust_method="ward.D2".
406
407 All statistical analyses were performed using R version 4.0.3.
408
409 **Results**
410 **Sample Characteristics**
411 The current study included 465,809 White individuals (53.9% female) with an average age of 56.6 yrs (sd
412 = 8.2 yrs) (Supplemental Table 1). Follow-up validation analyses were performed in 4,770 Blacks (60.2%
413 female) with an average age of 61.2 yrs (sd = 7.4 yrs). The majority of the data originated from the UKB
414 (93%). The bulk of the DNA used for mtDNA-CN estimation was derived from buffy coat (95.5%) while
415 the rest was derived from peripheral leukocytes (2.2%), whole blood (2.3%), or brain (< 0.2%). mtDNA-
416 CN estimated from Affymetrix genotyping arrays consisted of 97.9% of the data while the remainder
417 was derived from qPCR (1.8%) and WGS (0.3%).
418
419 **GWAS**
420 Previous work has demonstrated that the method used to measure mtDNA-CN can impact the strength
421 of association[36]. To account for potential differences across studies due to the different mtDNA-CN
422 measurements used, as well as the inclusion of blood cell counts as covariates in only a subset of the
423 cohorts, we took two approaches. First, we used a random effects model to perform meta-analyses,
424 allowing for different genetic effect size estimates across cohorts. Second, we
425 performed three complementary analyses in individuals who self-identified as White: 1) meta-analysis of
426 all available studies (n = 465,809); 2) meta-analysis of studies with available data for cell count
427 adjustment (n = 456,151); and 3) GWAS of UKB only (n = 440,266) (Figure 1). 77 loci were significant in

428 all three meta-analyses, and we identified 93 independent loci that were significant in at least one of the
429 analyses. In the meta-analysis of UKB-only data, 92 of the total 93 loci were identified (Supplemental
430 Figure 3). Given that >90% of the samples come from the UKB study, and the challenge of interpreting
431 effect size estimates from a random effects model, downstream analyses all use effect size estimates
432 from the UKB only analyses (Supplemental Table 6), which showed no evidence for population
433 substructure inflating test statistics, with a genomic inflation factor of 1.09 (Supplemental Figure 4). SNP
434 heritability estimated from BOLT-LMM[40] for mtDNA-CN adjusted for age and sex was 10.5% while
435 heritability for mtDNA-CN adjusted for age, sex, and cell counts was 7.4%, implying that some of the
436 mtDNA-CN heritability observed in previous studies could be due to heritability of cell type composition.
437 We also used SumHer[41] as an alternative approach to calculating SNP heritability, which returned a
438 comparable estimate of 7.0% for the cell-count corrected mtDNA-CN metric.
439 The most significant SNP associated with mtDNA-CN was a missense mutation in *LONP1* ($p = 3.00 \times 10^{-141}$),
440 a gene that encodes a mitochondrial protease that can directly bind mtDNA, and has been
441 shown to regulate *TFAM*, a transcription factor involved in mtDNA replication and repair (for review see
442 Gibellini *et al.*)[64].
443 Meta-analysis of the sex-stratified X chromosome results identified four loci significantly
444 associated with mtDNA-CN, with directionality consistent across the male and female stratified analyses
445 (Supplemental Table 7).
446
447 **Fine-mapping and Secondary Hits**
448 To identify additional independent SNPs within novel loci whose effects were masked by the original
449 significant SNP, as well as identify additional loci, we took two approaches. First, a conditional analysis
450 adjusting for the top 93 SNPs from the initial (primary) GWAS run revealed 3 novel loci and 19 additional
451 independent significant SNPs within existing loci. We also performed fine-mapping with susieR[42] and

452 discovered an additional 14 independent SNPs within existing loci. The majority of loci had only one 95%
453 credible set of SNPs; further, twenty of the credible sets contained only one SNP. However, many of the
454 credible sets contained greater than 50 SNPs after fine-mapping, and 12 of the 122 credible sets had a
455 missense SNP as the SNP with the highest PIP in the set. Using these two methods, we identified in total
456 129 independent SNPs across 96 autosomal loci (Supplemental Figure 5), while susieR fine-mapping and
457 conditional analyses for the X-chromosome loci did not reveal any additional secondary signals.
458
459 **Comparisons with Hägg et al. 2020**
460 Out of the 50 loci reported in Hägg et al. 2020[33], we replicate 38 loci in our cell-count adjusted analyses
461 (Supplemental Table 8). As the two GWASs both use UK Biobank data, this replication is unsurprising.
462 Out of the 12 loci that were not genome-wide significant in our cell-count adjusted analyses, 11 were
463 significant when we did not adjust our mtDNA-CN metric for cell counts, suggesting that cell-type
464 composition may be driving these signals. The current manuscript also reports 62 additional loci that are
465 not in the Hägg et al. 2020 study. This is likely due to increased power, as the sample size used for the
466 current analyses is nearly twice as large.
467
468 **Associations in Black Populations**
469 Examining the 129 autosomal SNPs from the Whites-only analysis, 99 were available in the Blacks-
470 only meta-analysis (n = 4770). After multiple testing correction, one of these SNPs was significant
471 (rs73349121, $p = 0.0001$), 9 were nominally significant ($p < 0.05$, with 5 expected), and 58/99 had a
472 direction of effect that was consistent with the Whites-only analyses (one-sided $p = 0.04$, Figure
473 2). Despite being under-powered, these results in the Blacks-only analyses provide evidence for similar
474 genetic effects in a different ancestry group.
475

476 **Gene Prioritization and Enrichment of mtDNA Depletion Syndrome Genes**

477 We integrated results from three different gene prioritization and functional annotation methods
478 (ANNOVAR[43], COLOC[46] , and DEPICT[48]) so that loci with nonsynonymous variants in gene exons were
479 prioritized first, with eQTL co-localization results considered second (Supplemental Table 9), and those
480 from DEPICT (Supplemental Table 10) were considered last (Supplemental Figure 2). For 20 loci, multiple
481 genes were assigned as analyses could not identify a single priority gene (Supplemental Table 11). As
482 eQTLGen did not evaluate X chromosome variants, three of the four X-chromosome loci were assigned
483 to the nearest gene. SLC25A5 was assigned to rs392020, as the second highest PIP SNP was an exonic
484 nonsynonymous variant.

485 We noted the identification of a number of mtDNA depletion syndrome genes in the priority list
486 and tested for enrichment of these known causal genes using a one-sided Fisher's exact test. For this
487 analysis, all genes for loci assigned to multiple genes were used, and genes for all primary and secondary
488 loci were considered. Our gene prioritization approach identified 7 of 16 mtDNA depletion genes
489 (Supplemental Table 12), consistent with a highly significant enrichment (one-sided $p = 3.09 \times 10^{-15}$).

490

491 **Gene Set Enrichment Analyses**

492 To avoid bias from a single locus with multiple functionally related genes contributing to a false-positive
493 signal, only one gene per unique locus was used, prioritizing genes assigned to primary loci. For loci with
494 multiple assigned genes, one gene was randomly selected for testing. To test for robustness of gene set
495 enrichment results, random selection was repeated 50 times, and only gene sets that were significantly
496 enriched for at least 40 iterations were retained. In all, a total of 100 genes were utilized for GO term
497 and KEGG pathway enrichment analyses. Using a Bonferroni-corrected p-value cutoff, 12 gene sets were
498 significantly enriched for all 50 iterations, including mitochondrial nucleoid, mitochondrial DNA

499 replication, and amyloid-beta clearance (Supplemental Table 13). No KEGG terms were significant across
500 multiple iterations.
501
502 **MetaXcan Gene Expression Analysis**
503 As a complementary approach to single-SNP analyses, we explored the associations between mtDNA-CN
504 and predicted gene expression using MetaXcan[51] MetaXcan incorporates multiple SNPs within a locus
505 along with a reference eQTL dataset to generate predicted gene expression levels. As our study
506 estimated mtDNA-CN derived from blood, we used whole blood gene expression eQTLs from the Gene-
507 Tissue Expression (GTEx) consortium[65] to predict gene expression in the UKB dataset. We identified
508 6,285 genes that had a predicted performance $p < 0.05$ (i.e., they had sufficient data to generate robust
509 gene expression levels) and were tested for association with mtDNA-CN. Of these genes, 74 were
510 significantly associated with mtDNA-CN ($p < 7.95 \times 10^{-6}$) (Figure 3, Supplemental Table 14), including 8
511 that were not identified through single-SNP analyses. Many of the significant genes have known
512 mitochondrial functions, notably the mtDNA transcription factor *TFAM* ($p = 1.09 \times 10^{-29}$) and
513 mitochondrial exonuclease *MGME1* ($p = 5.87 \times 10^{-23}$), genes known as causal for mtDNA depletion
514 syndromes[54,55]. Additionally, *LONP1*, *MRPL43*, and *BAK1*, are all genes with known mitochondrial
515 functions[66–68]. Bonferroni significant MetaXcan genes were used for gene enrichment analysis, finding
516 enrichment for "nucleobase metabolic process" ($p = 1.47 \times 10^{-4}$) and "mitochondrial fusion" ($p = 1.86 \times$
517 10^{-4}) (Supplemental Figure 6).
518
519 **PHEWAS-based SNP Clustering and Gene Set Enrichment**
520 mtDNA-CN is associated with numerous quantitative and qualitative phenotypes, many of which are
521 relevant to aging-related disease[3–5,9,10,13–16]. We hypothesized that this pleiotropy may reflect different
522 underlying functional domains captured by mtDNA-CN, and may be reflected in GWAS-identified SNPs

523 and their likely causal genes. To test this hypothesis, we used the UKB data to identify quantitative traits
524 associated with mtDNA-CN and selected 41 highly significant, non-redundant traits to test for
525 association with the mtDNA-CN GWAS SNPs (Supplemental Table 5, in PHEWAS = 1). We clustered SNPs
526 using the trait effect size (beta) divided by the mtDNA-CN effect size estimate, so that all effects are
527 standardized to the effect of the mtDNA-CN increasing allele for each locus. We identified 3 clusters of
528 SNPs (Supplemental Table 15, Figure 4A), with cluster 1 containing SNPs in which the mtDNA-CN
529 increasing allele is associated with decreased platelet count (PLT) (Figure 4B), increased mean platelet
530 volume (MPV) (Figure 4C), and platelet distribution width (PDW) (Figure 4D), consistent with a role in
531 platelet activation[69]. Cluster 2 is most strongly enriched for SNPs in which the mtDNA-CN increasing
532 allele is associated with increased PLT, plateletcrit (PCT, a measure of total platelet mass), serum
533 calcium (Figure 4E), serum phosphate, as well as decreased mean corpuscular volume (MCV) and mean
534 spherical cellular volume (Figure 4F) (Supplemental Table 16). The cluster 2 phenotypes implicate
535 megakaryocyte proliferation and proplatelet formation in addition to apoptosis and autophagy, and are
536 supported by the genes identified for this cluster (megakaryocyte proliferation and proplatelet
537 formation: *MYB*, *JAK2*[70], apoptosis and autophagy: *BAK1*, *BCL2, TYMP*)[71]. Gene set enrichment analysis
538 confirmed this, as cluster 2 genes are significantly enriched for extrinsic apoptosis signaling pathways in
539 the absence of ligand (Supplemental Table 17). Cluster 3 did not yield any specific trait enrichment (all
540 significant results reflected the strong enrichment observed in clusters 1-2); however, gene set
541 enrichment for this cluster identified multiple mtDNA-related gene ontology terms, including
542 mitochondrial DNA replication, gamma DNA polymerase complex, and mitochondrial nucleoid
543 (Supplemental Table 18).
544
545 **Determination of causal associations between mitochondrial function and mtDNA-CN associated traits**

546 The extensive pleiotropy and limited variance explained of nuclear DNA SNPs associated with mtDNA-CN
547 (<1% of the variance in mtDNA-CN explained by GWS loci when predicted into the ARIC cohort)
548 precludes the use of traditional Mendelian randomization (MR) approaches to establish causality
549 between mtDNA-CN and the 41 identified mtDNA-CN associated traits. As alternative approach, we
550 examined the association of mitochondrial SNPs with mtDNA-CN and the 41 traits, under the
551 assumption that these SNPs can only act through alteration of mitochondrial function, and thus a
552 significant association implies causality. Imputation and analyses of mitochondrial SNPs were run
553 stratified by genotyping array (see Methods), and then meta-analyzed using inverse-variance weighting.
554 After multi-test correction ($p < 9.5x10^{-6}$), we identified 45 SNPs associated with 1 or more of the traits,
555 ranging from 1 to 6 traits per SNP (Supplementary Table 19). To identify independent effects, we first
556 identified the most significantly associated trait for each SNP, highlighting 4 traits (aspartate
557 aminotransferase, creatinine, MCV, PCT) in which to run susieR to identify independent credible sets.
558 We identified 6 independent effects across the four traits, with MCV credible set 4 and platelet credible
559 set 1 representing the same effect. We note that 2 of the SNPs are in moderately high LD (MT73A_G and
560 MT7028C_T, r^2=0.67), however, conditional analyses demonstrate that MT73A_G is associated with
561 creatinine, and not MCV, and the reverse is seen for MT7028C_T (Supplemental Table 20). Leveraging
562 the haploid nature of the mitochondrial genome, we selected the directly genotyped SNP with the
563 highest PIP from each credible set (Supplemental Table 21), and identified 8 haplotypes with MAF >
564 0.005 (Supplemental Table 22). Comparing linear regression models with and without the haplotypes in
565 the model, we identify 14 traits nominally associated, and 9 traits significantly associated after
566 Bonferroni correction, with mtDNA genetic variation (Supplemental Table 23, Figure 5). These results
567 causally implicate mitochondrial function in a variety of cell related traits (MCV, MSCV, MPV, PCT,
568 Platelet), kidney function (creatinine), liver function (aspartate and alanine aminotransferases) and
569 mtDNA-CN.

570

571 **Association of mitochondrial function with mortality**

572 We have previously shown that mtDNA-CN is associated with overall mortality[9]. As above, we
573 collectively tested the mitochondrial haplotypes for association with mortality not due to external
574 causes (e.g., no accidents, falls, see Methods ; n = 24,622, median follow-up time = 4318 days), and
575 found a nominally significant association with overall mortality (p = 0.044, Figure 6). Given the
576 conflicting reports between increased mitochondrial function and both increased and decreased cancer
577 risk[16,72,73], we looked separately at cancer (n = 13,231) and non-cancer mortality (n = 11,391). While
578 there was no association with cancer mortality (p = 0.74), we saw a highly significant association with
579 non-cancer mortality ($p = 6.56 \times 10^{-4}$).

580

581 **Discussion**

582 We conducted a GWAS for mtDNA-CN using 465,809 individuals from the CHARGE consortium and the
583 UKB. We report 133 independent signals originating from 100 loci, the majority of which were not
584 identified in previous studies. Examining our GWS SNPs in a Black population, we observed a concordant
585 signal, suggesting that the genetic etiology of mtDNA-CN may be broadly similar across populations.
586 Using several functional follow-up methods, genes were assigned for each identified independent hit
587 and significant enrichment was observed for genes involved in mitochondrial DNA metabolism,
588 homeostasis, cell activation, and amyloid-beta clearance. In total, we assigned 128 unique genes to
589 independent GWAS signals associated with mtDNA-CN. We also identified 8 additional genes whose
590 predicted gene expression is associated with mtDNA-CN that could not be mapped back to GWS loci.
591 Finally, using a clustering approach based on SNP associations with various mtDNA-CN associated
592 phenotypes, we were able to functionally categorize SNPs, providing insight into biological pathways
593 that impact mtDNA-CN.

594 We note that during the preparation of this manuscript, a GWAS for mtDNA-CN performed in
595 295,150 unrelated individuals from the UK Biobank was published, which reported 50 genome-wide
596 significant regions[33]. Within our GWAS, we replicate 38 of these 50 genome-wide significant loci in our
597 cell count corrected analyses. An additional 11 out of the remaining 12 loci are genome-wide significant
598 when we do not adjust mtDNA-CN for cell count. While Hagg et. al adjust for cell type composition, this
599 difference suggests that their adjustment may not be fully capturing the effects of cell counts.
600 Additionally, our analyses report 59 additional loci that are not observed in the previous paper, largely
601 due to the increased power of our study.
602 We were able to identify a substantial proportion of the genes involved in mtDNA depletion
603 syndromes (7/16, $p = 3.09 \times 10^{-15}$ for enrichment), including *TWNK*, *TFAM*, *DGUOK*, *MGME1*, *RRM2B,*
604 *TYMP,* and *POLG*. mtDNA depletion syndromes can be broken down into 5 subtypes based on their
605 constellation of phenotypes[74] , and with the exception of cardiomyopathic subtypes (associated with
606 mutations in *AGK* and *SLC25A4*), we were able to identify at least 1 gene from the other 4 subtypes,
607 suggesting that our mtDNA-CN measurement in blood-derived DNA can identify genes widely relevant
608 to non-blood phenotypes. This finding is consistent with a large body of work showing that mtDNA-CN
609 measured in blood is associated with numerous aging-related phenotypes for which the primary tissue
610 of interest is not blood (e.g. chronic kidney disease[13], heart failure[11], and diabetes[75]). Also consistent
611 with this finding is recent work demonstrating that mtDNA-CN measured in blood is associated with
612 mtRNA expression across numerous non-blood tissues, suggesting a link between mitochondrial
613 function measured in blood and other tissues[76].
614 In addition to identifying the mtDNA depletion syndrome genes directly linked to mitochondrial
615 DNA metabolic processes, DNA replication, and genome maintenance, we also identify genes which play
616 a role in mitochondrial function. The top GWAS hit is a missense mutation in *LONP1*, which encodes a
617 mitochondrial protease that has been shown to cause mitochondrial cytopathy and reduced respiratory

618 chain activity[77,78]. Interestingly, this missense mutation was recently found to be associated with
619 mitochondrial tRNA methylation levels[79]. Additional genes known to impact mitochondrial function
620 include *MFN1*, which encodes a mediator of mitochondrial fusion[80,81], *STMP1*, which plays a role in
621 mitochondrial respiration[82], and *MRPS35*, which encodes a ribosomal protein involved in protein
622 synthesis in the mitochondrion[83,84].

623 Using a combination of gene-based tests and gene prioritization using functional annotation,
624 pathway analyses reveal enrichment for numerous mitochondrial related pathways, as well as those
625 involved in regulation of cell differentiation ($p < 1.08 \times 10^{-5}$), homeostatic processes ($p < 3.77 \times 10^{-6}$), and
626 cellular response to stress ($p < 3.49 \times 10^{-6}$) (Supplemental Table 13). These results provide additional
627 evidence for the broad role played by mitochondria in numerous aspects of cellular function. Of
628 particular interest, the GO term for amyloid beta is significantly enriched, reinforcing a link between
629 mtDNA-CN and neurodegenerative disease[85–87]. Previous work from our lab using the UKB has shown
630 that increased mtDNA-CN is associated with lower rates of prevalent neurodegenerative disease, and is
631 predictive for decreased risk of incident neurodegenerative disease[76]. mtDNA-CN is also known to be
632 decreased in the frontal cortex of Alzheimer's disease (AD) patients[88]. Interestingly, the four GWAS-
633 identified genes driving the enrichment for amyloid-beta clearance are all related to regulation of lipid
634 levels, and lipid homeostasis within the brain is known to play an important role in Alzheimer's
635 disease[89]. *APOE*, one of the most well-known risk genes for Alzheimer's disease, is a cholesterol carrier
636 involved in lipid transport, and the ApoE-ε4 isoform involved in AD pathogenesis is associated with
637 mitochondrial dysfunction and oxidative distress in the human brain[90]; *CD36* is a platelet glycoprotein
638 which mediates the response to amyloid-beta accumulation[91]; *LDLR* is a low-density lipoprotein receptor
639 associated with AD[92]; and *ABCA7* is a phospholipid transporter[93]. *ABCA7* loss of function variants are
640 enriched in both AD and Parkinson's disease (PD) patients[94], suggesting a broad role across
641 neurodegenerative diseases.

642 Given the integral role of mitochondria in cellular function, from ATP formation and energy
643 production, signaling through reactive oxygen species, and apoptosis mediation, there is a strong basis
644 to *a priori* assume that genetic variants associated with mtDNA-CN are likely to be highly pleiotropic.
645 Indeed, mtDNA-CN itself is associated with numerous phenotypes (Supplemental Table 5). Through our
646 PHEWAS-based clustering approach using 41 mtDNA-CN associated phenotypes, we uncovered
647 phenotypic associations between three distinct clusters of GWS mtDNA-CN associated SNPs. Cluster 1
648 was characterized by increased MPV, PDW, and decreased PLT (note that measured MPV and PLT are
649 generally inversely correlated to maintain hemostasis), which are the hallmarks of platelet activation[69].
650 The link between platelets and mtDNA-CN has typically revolved around platelet count, as platelets have
651 functional mitochondria, but do not have a nucleus. Given that the mtDNA-CN measurement is the ratio
652 between mtDNA and nuclear DNA, increased platelets, all else being equal, would directly equate with
653 increased mtDNA-CN. We note that the mtDNA-CN metric used in this GWAS was adjusted for platelet
654 count, likely increasing the ability to detect variants that impact mtDNA-CN through increased platelet
655 activation. Examining the genes within this cluster suggests roles for actin formation and regulation
656 (*TPM4*, *PACSIN2*)[95,96] and vesicular transport and endocytic trafficking (*DNM3*, *EHD3*)[97,98] in platelet
657 activation.
658 Cluster 2 is most strongly enriched for SNPs in which the mtDNA-CN increasing allele is
659 associated with increased PLT/PCT and serum calcium/phosphate. Examining the genes assigned to the
660 cluster, we implicate megakaryocyte proliferation and proplatelet formation (*MYB*, *JAK2*)[70], and
661 apoptosis and autophagy (*BAK1*, *BCL2, TYMP*)[71]. Megakaryocytes are used to form proplatelets, and the
662 process includes an important role for both intra- and extracellular calcium levels[99]. A role for apoptosis,
663 and specifically *BCL2*, in proplatelet formation and platelet release has been suggested[100,101], however
664 work in mice has suggested that apoptosis does not play a direct role in these processes[102].
665 Nevertheless, apoptosis is important for platelet lifespan[103].

666 Cluster 3 was particularly challenging to interpret, given that no particular phenotype was
667 enriched relative to the non-cluster 3 SNPs. We note that this cluster appeared to be enriched for the
668 mtDNA depletion syndrome genes, containing 6/7 genes identified in the GWAS, and significantly
669 enriched for GO Terms related mitochondrial DNA. Additionally, genes in cluster 3 were significantly
670 enriched for low-density lipoprotein particle binding, suggesting a role for lipid homeostasis. Closer
671 inspection of cluster 3 genes reveals a number of genes known to be associated with lipid levels (*LIPC*,
672 *CETP*, *LDLR*, *APOE*). While lipids play a role in both energy metabolism (largely through fatty acids) and
673 cellular membrane formation, a link to mtDNA-CN and/or mitochondrial function is not well-established.
674 A strong rationale for the study of mtDNA-CN is the underlying assumption that it reflects
675 mitochondrial function and is readily measured, often from existing data. A serious complication to the
676 interpretation of the role of mitochondrial function in various traits has been use of blood-derived
677 measurements, which can be confounded by differences in cell counts across individuals. Mendelian
678 randomization has been widely used to infer causality between traits (e.g. LDL and CAD)[104], but is only
679 robust under conditions of little to no pleiotropy[105] and its power is a function of variance explained. For
680 mtDNA-CN, the extensive pleiotropy and small amount of variance explained of GWS variants (<1%)
681 prevents the use of traditional MR approaches. As an alternative approach, we analyzed associations
682 between mitochondrial DNA variants and mtDNA-CN associated phenotypes. Presumably, variants
683 located on the mitochondrial genome are only able to modify phenotypes through modulating
684 mitochondrial function, allowing for causal inference. Our analyses revealed significant relationships
685 between mitochondrial variants and creatinine, aspartate aminotransferase, MCV, and PCT. Creatinine
686 and aspartate aminotransferase are markers of kidney and liver function respectively, and supporting
687 these findings, mtDNA-CN has been linked to both chronic kidney disease[13] and non-alcoholic fatty liver
688 disease[106]. We also find a highly significant association between mitochondrial variation and non-cancer

689 mortality, adding evidence for a causal relationship to previous findings showing mtDNA-CN is
690 associated with all-cause mortality[9].
691 Several limitations should be noted. First, despite the large sample size and numerous loci
692 identified, we are likely missing a great deal of the true signal, as our SNP heritability estimates through
693 SumHer and BOLT-LMM were 7.0% and 7.4% respectively, while previous studies have estimated
694 mtDNA-CN heritability to be 65%[23]. Finally, while we have adjusted our mtDNA-CN metric for a variety of
695 confounders, it is important to note that mtDNA-CN can be influenced by a variety of environmental
696 factors including smoking[107] and drugs, which have not been adjusted for in these analyses.
697 In summary, we performed the largest-to-date GWAS for mtDNA-CN, including almost 500,000
698 individuals. We identified three distinct groups of SNPs associated with mtDNA-CN that are related to
699 platelet activation, megakaryocyte formation and apoptotic processes, and showed clear enrichment for
700 genes involved in mtDNA depletion and nucleotide regulation. Additionally, we find that mitochondrial
701 variants are significantly associated with creatinine, aspartate aminotransferase, MCV, and PCT,
702 implying a causal relationship between mitochondrial function and these phenotypes. Finally, we
703 provide strong evidence that mitochondrial function is causal for non-cancer mortality. Given the role of
704 mtDNA-CN, and, by proxy, mitochondrial function in aging-related disease, this work begins to unravel
705 the many varied underlying mechanisms through which mitochondrial function impacts human health.
706
707 **Supplemental Data**
708 Supplemental Data include six supplemental figures, twenty-two supplemental tables, and supplemental
709 methods.
710
711 **Declaration of Interests**

712 Psaty serves on the Steering Committee of the Yale Open Data Access Project funded by Johnson &
713 Johnson. All other authors declare no competing interests.
714
715 **Acknowledgements**
716 This work was supported by National Heart, Lung and Blood Institute, National Institutes of Health (NIH)
717 grants R01HL13573 and R01HL144569 (RJL, SYY, CAC, DEA), NIH grant P01-AG027734 (GA, YK, NB, AB)
718 and the National Center for Advancing Translational Sciences, NIH, through Grant KL2TR002490 (LMR).
719 The content is solely the responsibility of the authors and does not necessarily represent the official
720 views of the NIH. LMR was also funded by T32 HL129982.
721
722 This research was conducted using data from the Genotype-Tissue Expression (GTEx) project (dbGaP
723 accession: phs000424.**v8**.p2). The GTEx project was supported by the Common Fund of the Office of the
724 Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.
725
726 This research was also conducted using the UK Biobank Resource under Application Number 17731.

727 The Atherosclerosis Risk in Communities study (dbGaP accession: phs000280.**v7**.p1) has been funded in
728 whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National
729 Institutes of Health, Department of Health and Human Services (contract numbers HSN268201700001I,
730 HHSN268201700002I, HHSN268201700003I, HHSN268201700004I and HHSN268201700005I),
731 R01HL087641, R01HL059367 and R01HL086694; National Human Genome Research Institute contract
732 U01HG004402; and National Institutes of Health contract HHSN268200625226C. Funding support for
733 "Building on GWAS for NHLBI diseases: the U.S. CHARGE consortium" was provided by the NIH through
734 the American Recovery and Reinvestment Act of 2009 (ARRA) (5RC2HL102419). Sequencing was carried
735 out at the Baylor College of Medicine Human Genome Sequencing Center and supported by the National

736 Human Genome Research Institute grants U54 HG003273 and UM1 HG008898. The authors thank the
737 staff and participants of the ARIC study for their important contributions. Infrastructure was partly
738 supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH
739 Roadmap for Medical Research.

740 This CHS research was supported by NHLBI contracts HHSN268201200036C,
741 HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081,
742 N01HC85082, N01HC85083, N01HC85086, 75N92021D00006; and NHLBI grants U01HL080295,
743 R01HL087652, R01HL105756, R01HL103612, R01HL120393, and U01HL130114 with additional
744 contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional
745 support was provided through R01AG023629 from the National Institute on Aging (NIA). A full list of
746 principal CHS investigators and institutions can be found at CHS-NHLBI.org. The provision of genotyping
747 data was supported in part by the National Center for Advancing Translational Sciences, CTSI
748 grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes
749 Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research
750 Center.

751 The FHS phenotype-genotype analyses were supported by the National Institute of Aging
752 (U34AG051418). This research was conducted in part using data and resources from the Framingham
753 Heart Study of the National Heart Lung and Blood Institute of the National Institutes of Health and
754 Boston University School of Medicine. This work was partially supported by the National Heart, Lung and
755 Blood Institute's Framingham Heart Study (Contract No. N01-HC-25195, HHSN268201500001) and its
756 contract with Affymetrix, Inc for genotyping services (Contract No. N02-HL-6-4278). Genotyping, quality
757 control and calling of the Illumina HumanExome BeadChip in the Framingham Heart Study was
758 supported by funding from the National Heart, Lung and Blood Institute Division of Intramural Research
759 (Daniel Levy and Christopher J. O'Donnell, Principle Investigators). The authors thank the participants for
760 their dedication to the study. The authors are pleased to acknowledge that the computational work
761 reported on in this paper was performed on the Shared Computing Cluster which is administered by
762 Boston University's Research Computing Services. URL: www.bu.edu/tech/support/research/.
763
764 MESA and the MESA SHARe projects (dbGaP accession: phs000209.**v13**.p3) are conducted and
765 supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA
766 investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-
767 HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003,
768 N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007,
769 N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-
770 001079, and UL1-TR-001420. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-
771 64278. Genotyping was performed at Affymetrix (Santa Clara, California, USA) and the Broad Institute of
772 Harvard and MIT (Boston, Massachusetts, USA) using the Affymetrix Genome-Wide Human SNP Array
773 6.0. The provision of genotyping data was supported in part by the National Center for Advancing
774 Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive

775 and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes
776 Endocrinology Research Center.
777
778 ROS/MAP is supported by the Translational Genomics Research Institute and National Institute on Aging
779 (NIA) through grants U01AG46152, U01AG61256, P30AG10161, R01AG17917, RF1AG15819,
780 R01AG30146.
781
782 SHIP and SHIP-TREND are part of the Community Medicine Research net of the University of Greifswald,
783 Germany, which is funded by the Federal Ministry of Education and Research (grants no. 01ZZ9603,
784 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State
785 of Mecklenburg-West Pomerania, and the network 'Greifswald Approach to Individualized Medicine
786 (GANI_MED)' funded by the Federal Ministry of Education and Research (grant 03IS2061A).

787 This research was funded in whole, or in part, by the Wellcome Trust [Grant ref: 217065/Z/19/Z]. For the
788 purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted
789 Manuscript version arising from this submission.

790 **Web Resources**

791 REVIGO was accessed at http://revigo.irb.hr/.

792

793 **Data and Code Availability**

794 All data used in this manuscript is available through either the UKBiobank and CHARGE consortiums.
795 Code and scripts are available in a zipped file at https://www.arkinglab.org/resources/.

796

797

798 **References**

799 1. Wallace, D.C. (1992). Diseases of the Mitochondrial Dna. Annual Review of Biochemistry *61*, 1175–
800 1212.

801 2. Vakifahmetoglu-Norberg, H., Ouchida, A.T., and Norberg, E. (2017). The role of mitochondria in
802 metabolism and cell death. Biochem. Biophys. Res. Commun. *482*, 426–431.

803 3. Dai, D.-F., Rabinovitch, P.S., and Ungvari, Z. (2012). Mitochondria and cardiovascular aging. Circ. Res.
804 *110*, 1109–1124.

805 4. Cui, H., Kong, Y., and Zhang, H. (2012). Oxidative stress, mitochondrial dysfunction, and aging. J Signal
806 Transduct *2012*, 646354.

807 5. Herst, P.M., Rowe, M.R., Carson, G.M., and Berridge, M.V. (2017). Functional Mitochondria in Health
808 and Disease. Front Endocrinol (Lausanne) *8*, 296.

809 6. Liu, C.-S., Tsai, C.-S., Kuo, C.-L., Chen, H.-W., Lii, C.-K., Ma, Y.-S., and Wei, Y.-H. (2003). Oxidative stress-
810 related alteration of the copy number of mitochondrial DNA in human leukocytes. Free Radic Res *37*,
811 1307–1317.

812 7. Guha, M., and Avadhani, N.G. (2013). Mitochondrial retrograde signaling at the crossroads of tumor
813 bioenergetics, genetics and epigenetics. Mitochondrion *13*, 577–591.

814 8. Malik, A.N., and Czajka, A. (2013). Is mitochondrial DNA content a potential biomarker of
815 mitochondrial dysfunction? Mitochondrion *13*, 481–492.

816 9. Ashar, F.N., Moes, A., Moore, A.Z., Grove, M.L., Chaves, P.H.M., Coresh, J., Newman, A.B., Matteini,
817 A.M., Bandeen-Roche, K., Boerwinkle, E., et al. (2015). Association of mitochondrial DNA levels with
818 frailty and all-cause mortality. J. Mol. Med. *93*, 177–186.

819 10. Ashar, F.N., Zhang, Y., Longchamps, R.J., Lane, J., Moes, A., Grove, M.L., Mychaleckyj, J.C., Taylor,
820 K.D., Coresh, J., Rotter, J.I., et al. (2017). Association of Mitochondrial DNA Copy Number With
821 Cardiovascular Disease. JAMA Cardiol *2*, 1247–1255.

822 11. Hong, Y.S., Longchamps, R.J., Zhao, D., Castellani, C.A., Loehr, L.R., Chang, P.P., Matsushita, K., Grove,
823 M.L., Boerwinkle, E., Arking, D.E., et al. (2020). Mitochondrial DNA Copy Number and Incident Heart
824 Failure: The Atherosclerosis Risk in Communities (ARIC) Study. Circulation *141*, 1823–1825.

825 12. Zhao, D., Bartz, T.M., Sotoodehnia, N., Post, W.S., Heckbert, S.R., Alonso, A., Longchamps, R.J.,
826 Castellani, C.A., Hong, Y.S., Rotter, J.I., et al. (2020). Mitochondrial DNA copy number and incident atrial
827 fibrillation. BMC Medicine *18*, 246.

828 13. Tin, A., Grams, M.E., Ashar, F.N., Lane, J.A., Rosenberg, A.Z., Grove, M.L., Boerwinkle, E., Selvin, E.,
829 Coresh, J., Pankratz, N., et al. (2016). Association between Mitochondrial DNA Copy Number in
830 Peripheral Blood and Incident CKD in the Atherosclerosis Risk in Communities Study. J Am Soc Nephrol
831 *27*, 2467–2473.

832 14. Pyle, A., Anugrha, H., Kurzawa-Akanbi, M., Yarnall, A., Burn, D., and Hudson, G. (2016). Reduced
833 mitochondrial DNA copy number is a biomarker of Parkinson's disease. Neurobiol. Aging *38*, 216.e7-
834 216.e10.

835 15. Wei, W., Keogh, M.J., Wilson, I., Coxhead, J., Ryan, S., Rollinson, S., Griffin, H., Kurzawa-Akanbi, M.,
836 Santibanez-Koref, M., Talbot, K., et al. (2017). Mitochondrial DNA point mutations and relative copy
837 number in 1363 disease and control human brains. Acta Neuropathologica Communications *5*, 13.

838 16. Reznik, E., Miller, M.L., Şenbabaoğlu, Y., Riaz, N., Sarungbam, J., Tickoo, S.K., Al-Ahmadie, H.A., Lee,
839 W., Seshan, V.E., Hakimi, A.A., et al. (2016). Mitochondrial DNA copy number variation across human
840 cancers. ELife *5*, e10769.

841 17. Hurtado-Roca, Y., Ledesma, M., Gonzalez-Lazaro, M., Moreno-Loshuertos, R., Fernandez-Silva, P.,
842 Enriquez, J.A., and Laclaustra, M. (2016). Adjusting MtDNA Quantification in Whole Blood for Peripheral
843 Blood Platelet and Leukocyte Counts. PLOS ONE *11*, e0163770.

844 18. Knez, J., Winckelmans, E., Plusquin, M., Thijs, L., Cauwenberghs, N., Gu, Y., Staessen, J.A., Nawrot,
845 T.S., and Kuznetsova, T. (2016). Correlates of Peripheral Blood Mitochondrial DNA Content in a General
846 Population. Am J Epidemiol *183*, 138–146.

847 19. Kumar, P., Efstathopoulos, P., Millischer, V., Olsson, E., Wei, Y.B., Brüstle, O., Schalling, M.,
848 Villaescusa, J.C., Ösby, U., and Lavebratt, C. (2018). Mitochondrial DNA copy number is associated with
849 psychosis severity and anti-psychotic treatment. Sci Rep *8*, 12743.

850 20. Urata, M., Koga-Wada, Y., Kayamori, Y., and Kang, D. (2008). Platelet contamination causes large
851 variation as well as overestimation of mitochondrial DNA content of peripheral blood mononuclear cells.
852 Ann Clin Biochem *45*, 513–514.

853 21. Clay Montier, L.L., Deng, J.J., and Bai, Y. (2009). Number matters: control of mammalian
854 mitochondrial DNA copy number. J Genet Genomics *36*, 125–131.

855 22. Tang, Y., Schon, E.A., Wilichowski, E., Vazquez-Memije, M.E., Davidson, E., and King, M.P. (2000).
856 Rearrangements of Human Mitochondrial DNA (mtDNA): New Insights into the Regulation of mtDNA
857 Copy Number and Gene Expression. Mol Biol Cell *11*, 1471–1485.

858 23. Xing, J., Chen, M., Wood, C.G., Lin, J., Spitz, M.R., Ma, J., Amos, C.I., Shields, P.G., Benowitz, N.L., Gu,
859 J., et al. (2008). Mitochondrial DNA content: its genetic heritability and association with renal cell
860 carcinoma. J. Natl. Cancer Inst. *100*, 1104–1112.

861 24. Carling, P.J., Cree, L.M., and Chinnery, P.F. (2011). The implications of mitochondrial DNA copy
862 number regulation during embryogenesis. Mitochondrion *11*, 686–692.

863 25. Harvey, A., Gibson, T., Lonergan, T., and Brenner, C. (2011). Dynamic regulation of mitochondrial
864 function in preimplantation embryos and embryonic stem cells. Mitochondrion *11*, 829–838.

865 26. Copeland, W.C. (2014). Defects of Mitochondrial DNA Replication. J Child Neurol *29*, 1216–1224.

866 27. Mandel, H., Szargel, R., Labay, V., Elpeleg, O., Saada, A., Shalata, A., Anbinder, Y., Berkowitz, D.,
867 Hartman, C., Barak, M., et al. (2001). The deoxyguanosine kinase gene is mutated in individuals with
868 depleted hepatocerebral mitochondrial DNA. Nat. Genet. *29*, 337–341.

869 28. Wang, L., Limongelli, A., Vila, M.R., Carrara, F., Zeviani, M., and Eriksson, S. (2005). Molecular insight
870 into mitochondrial DNA depletion syndrome in two patients with novel mutations in the
871 deoxyguanosine kinase and thymidine kinase 2 genes. Mol. Genet. Metab. *84*, 75–82.

872 29. Rusecka, J., Kaliszewska, M., Bartnik, E., and Tońska, K. (2018). Nuclear genes involved in
873 mitochondrial diseases caused by instability of mitochondrial DNA. J Appl Genet *59*, 43–57.

874 30. Cai, N., Li, Y., Chang, S., Liang, J., Lin, C., Zhang, X., Liang, L., Hu, J., Chan, W., Kendler, K.S., et al.
875 (2015). Genetic Control over mtDNA and Its Relationship to Major Depressive Disorder. Curr Biol *25*,
876 3170–3177.

877 31. Workalemahu, T., Enquobahrie, D.A., Tadesse, M.G., Hevner, K., Gelaye, B., Sanchez, S., and
878 Williams, M.A. (2017). Genetic Variations Related to Maternal Whole Blood Mitochondrial DNA Copy
879 Number: A Genome-Wide and Candidate Gene Study. J Matern Fetal Neonatal Med *30*, 2433–2439.

880 32. Guyatt, A.L., Brennan, R.R., Burrows, K., Guthrie, P.A.I., Ascione, R., Ring, S.M., Gaunt, T.R., Pyle, A.,
881 Cordell, H.J., Lawlor, D.A., et al. (2019). A genome-wide association study of mitochondrial DNA copy
882 number in two population-based cohorts. Human Genomics *13*, 6.

883 33. Hägg, S., Jylhävä, J., Wang, Y., Czene, K., and Grassmann, F. (2020). Deciphering the genetic and
884 epidemiological landscape of mitochondrial DNA abundance. Hum Genet.

885 34. Psaty Bruce M., O'Donnell Christopher J., Gudnason Vilmundur, Lunetta Kathryn L., Folsom Aaron R.,
886 Rotter Jerome I., Uitterlinden André G., Harris Tamara B., Witteman Jacqueline C.M., and Boerwinkle
887 Eric (2009). Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium.
888 Circulation: Cardiovascular Genetics *2*, 73–80.

889 35. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D.,
890 Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic
891 data. Nature *562*, 203–209.

892 36. Longchamps, R.J., Castellani, C.A., Yang, S.Y., Newcomb, C.E., Sumpter, J.A., Lane, J., Grove, M.L.,
893 Guallar, E., Pankratz, N., Taylor, K.D., et al. (2020). Evaluation of mitochondrial DNA copy number
894 estimation techniques. PLOS ONE *15*, e0228166.

895 37. MitoPipeline: Generating Mitochondrial copy number estimates from SNP array data in Genvisis.

896 38. Han, B., and Eskin, E. (2011). Random-Effects Model Aimed at Discovering Associations in Meta-
897 Analysis of Genome-wide Association Studies. The American Journal of Human Genetics *88*, 586–598.

898 39. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide
899 association scans. Bioinformatics *26*, 2190–2191.

900 40. Loh, P.-R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman,
901 D.I., Ridker, P.M., Neale, B.M., Berger, B., et al. (2015). Efficient Bayesian mixed model analysis increases
902 association power in large cohorts. Nat Genet *47*, 284–290.

903 41. Speed, D. (2019). SumHer better estimates the SNP heritability of complex traits from summary
904 statistics. Nature Genetics *51*, 12.

905 42. Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable
906 selection in regression, with application to genetic fine mapping. Journal of the Royal Statistical Society:
907 Series B (Statistical Methodology) *82*, 1273–1300.

908 43. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants
909 from high-throughput sequencing data. Nucleic Acids Research *38*, e164–e164.

910 44. Sherry, S.T., Ward, M., and Sirotkin, K. (1999). dbSNP-database for single nucleotide polymorphisms
911 and other classes of minor genetic variation. Genome Res *9*, 677–679.

912 45. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B.,
913 Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current
914 status, taxonomic expansion, and functional annotation. Nucleic Acids Res *44*, D733-745.

915 46. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V.
916 (2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary
917 Statistics. PLOS Genetics *10*, e1004383.

918 47. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A.,
919 Kreuzhuber, R., Kasela, S., et al. (2018). Unraveling the polygenic architecture of complex traits using
920 blood eQTL metaanalysis. BioRxiv 447367.

921 48. Pers, T.H., Karjalainen, J.M., Chan, Y., Westra, H.-J., Wood, A.R., Yang, J., Lui, J.C., Vedantam, S.,
922 Gustafsson, S., Esko, T., et al. (2015). Biological interpretation of genome-wide association studies using
923 predicted gene functions. Nature Communications *6*, 5890.

924 49. Smyth, G., Hu, Y., Ritchie, M., Silver, J., Wettenhall, J., McCarthy, D., Wu, D., Shi, W., Phipson, B., Lun,
925 A., et al. (2021). limma: Linear Models for Microarray Data (Bioconductor version: Release (3.12)).

926 50. Young, M.D., Wakefield, M.J., Smyth, G.K., and Oshlack, A. (2010). Gene ontology analysis for RNA-
927 seq: accounting for selection bias. Genome Biology *11*, R14.

928 51. Barbeira, A.N., Dickinson, S.P., Bonazzola, R., Zheng, J., Wheeler, H.E., Torres, J.M., Torstenson, E.S.,
929 Shah, K.P., Garcia, T., Edwards, T.L., et al. (2018). Exploring the phenotypic consequences of tissue
930 specific gene expression variation inferred from GWAS summary statistics. Nat Commun *9*, 1825.

931 52. Barbeira, A.N., Pividori, M., Zheng, J., Wheeler, H.E., Nicolae, D.L., and Im, H.K. (2019). Integrating
932 predicted transcriptome from multiple tissues improves association detection. PLOS Genetics *15*,
933 e1007889.

934 53. Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists
935 of gene ontology terms. PLoS One *6*, e21800.

936 54. Stiles, A.R., Simon, M.T., Stover, A., Eftekharian, S., Khanlou, N., Wang, H.L., Magaki, S., Lee, H.,
937 Partynski, K., Dorrani, N., et al. (2016). Mutations in TFAM, encoding mitochondrial transcription factor
938 A, cause neonatal liver failure associated with mtDNA depletion. Mol Genet Metab *119*, 91–99.

939 55. Kornblum, C., Nicholls, T.J., Haack, T.B., Schöler, S., Peeva, V., Danhauser, K., Hallmann, K., Zsurka, G.,
940 Rorbach, J., Iuso, A., et al. (2013). Loss-of-function mutations in MGME1 impair mtDNA replication and
941 cause multisystemic mitochondrial disease. Nat Genet *45*, 214–219.

942 56. El-Hattab, A.W., and Scaglia, F. (2013). Mitochondrial DNA depletion syndromes: review and updates
943 of genetic basis, manifestations, and therapeutic options. Neurotherapeutics *10*, 186–198.

944 57. Millard, L.A.C., Davies, N.M., Gaunt, T.R., Davey Smith, G., and Tilling, K. (2018). Software Application
945 Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. Int J Epidemiol *47*,
946 29–35.

947 58. Konopka, T. (2020). umap: Uniform Manifold Approximation and Projection.

948 59. Hahsler, M., Piekenbrock, M., Arya, S., and Mount, D. (2019). dbscan: Density Based Clustering of
949 Applications with Noise (DBSCAN) and Related Algorithms.

950 60. Delaneau, O., Zagury, J.-F., Robinson, M.R., Marchini, J.L., and Dermitzakis, E.T. (2019). Accurate,
951 scalable and integrative haplotype estimation. Nat Commun *10*,.

952 61. Rubinacci, S., Delaneau, O., and Marchini, J. (2020). Genotype imputation using the Positional
953 Burrows Wheeler Transform. PLOS Genetics *16*, e1009049.

954 62. Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A.,
955 Clark, A.G., Donnelly, P., Eichler, E.E., et al. (2015). A global reference for human genetic variation.
956 Nature *526*, 68–74.

957 63. Shen, L., Attimonelli, M., Bai, R., Lott, M.T., Wallace, D.C., Falk, M.J., and Gai, X. (2018). MSeqDR
958 mvTool: A mitochondrial DNA Web and API resource for comprehensive variant annotation, universal
959 nomenclature collation, and reference genome conversion. Hum Mutat *39*, 806–810.

960 64. Gibellini, L., De Gaetano, A., Mandrioli, M., Van Tongeren, E., Bortolotti, C.A., Cossarizza, A., and
961 Pinti, M. (2020). The biology of Lonp1: More than a mitochondrial protease. Int Rev Cell Mol Biol *354*, 1–
962 61.

963 65. GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet *45*, 580–585.

964 66. Liu, T., Lu, B., Lee, I., Ondrovicová, G., Kutejová, E., and Suzuki, C.K. (2004). DNA and RNA binding by
965 the mitochondrial lon protease is regulated by nucleotide and protein substrate. J Biol Chem *279*,
966 13902–13910.

967 67. Sharma, M.R., Koc, E.C., Datta, P.P., Booth, T.M., Spremulli, L.L., and Agrawal, R.K. (2003). Structure
968 of the mammalian mitochondrial ribosome reveals an expanded functional role for its component
969 proteins. Cell *115*, 97–108.

970 68. Shimizu, S., Narita, M., and Tsujimoto, Y. (1999). Bcl-2 family proteins regulate the release of
971 apoptogenic cytochrome c by the mitochondrial channel VDAC. Nature *399*, 483–487.

972 69. Vagdatli, E., Gounari, E., Lazaridou, E., Katsibourlia, E., Tsikopoulou, F., and Labrianou, I. (2010).
973 Platelet distribution width: a simple, practical and specific marker of activation of coagulation.
974 Hippokratia *14*, 28–32.

975 70. PathCards :: Factors involved in megakaryocyte development and platelet production Pathway and
976 related pathways.

977 71. PathCards :: Apoptosis and Autophagy Pathway and related pathways.

978 72. Yuan, Y., Ju, Y.S., Kim, Y., Li, J., Wang, Y., Yoon, C.J., Yang, Y., Martincorena, I., Creighton, C.J.,
979 Weinstein, J.N., et al. (2020). Comprehensive molecular characterization of mitochondrial genomes in
980 human cancers. Nature Genetics *52*, 342–352.

981 73. Mizumachi, T., Muskhelishvili, L., Naito, A., Furusawa, J., Fan, C.-Y., Siegel, E.R., Kadlubar, F.F., Kumar,
982 U., and Higuchi, M. (2008). Increased Distributional Variance of Mitochondrial DNA Content Associated
983 With Prostate Cancer Cells as Compared With Normal Prostate Cells. Prostate *68*, 408–417.

984 74. Basel, D. (2020). Mitochondrial DNA Depletion Syndromes. Clin Perinatol *47*, 123–141.

985 75. DeBarmore, B., Longchamps, R.J., Zhang, Y., Kalyani, R.R., Guallar, E., Arking, D.E., Selvin, E., and
986 Young, J.H. (2020). Mitochondrial DNA copy number and diabetes: the Atherosclerosis Risk in
987 Communities (ARIC) study. BMJ Open Diabetes Res Care *8*,.

988 76. Yang, S.Y., Castellani, C.A., Longchamps, R.J., Pillalamarri, V.K., O'Rourke, B., Guallar, E., and Arking,
989 D.E. (2021). Blood-derived mitochondrial DNA copy number is associated with gene expression across
990 multiple tissues and is predictive for incident neurodegenerative disease. Genome Res.

991 77. Hannah-Shmouni, F., MacNeil, L., Brady, L., Nilsson, M.I., and Tarnopolsky, M. (2019). Expanding the
992 Clinical Spectrum of LONP1-Related Mitochondrial Cytopathy. Front Neurol *10*, 981.

993 78. Grainha, T.R.R., Jorge, P.A. da S., Pérez-Pérez, M., Pérez Rodríguez, G., Pereira, M.O.B.O., and
994 Lourenço, A.M.G. (2018). Exploring anti-quorum sensing and anti-virulence based strategies to fight
995 Candida albicans infections: an in silico approach. FEMS Yeast Res *18*,.

996 79. Ali, A.T., Idaghdour, Y., and Hodgkinson, A. (2020). Analysis of mitochondrial m1A/G RNA
997 modification reveals links to nuclear genetic variants and associated disease processes. Communications
998 Biology *3*, 1–11.

999 80. Schrepfer, E., and Scorrano, L. (2016). Mitofusins, from Mitochondria to Metabolism. Mol Cell *61*,
1000 683–694.

1001 81. Ishihara, N., Eura, Y., and Mihara, K. (2004). Mitofusin 1 and 2 play distinct roles in mitochondrial
1002 fusion reactions via GTPase activity. J Cell Sci *117*, 6535–6546.

1003 82. Zhang, D., Xi, Y., Coccimiglio, M.L., Mennigen, J.A., Jonz, M.G., Ekker, M., and Trudeau, V.L. (2012).
1004 Functional prediction and physiological characterization of a novel short trans-membrane protein 1 as a
1005 subunit of mitochondrial respiratory complexes. Physiol Genomics *44*, 1133–1140.

1006 83. Cavdar Koc, E., Burkhart, W., Blackburn, K., Moseley, A., and Spremulli, L.L. (2001). The small subunit
1007 of the mammalian mitochondrial ribosome. Identification of the full complement of ribosomal proteins
1008 present. J Biol Chem *276*, 19363–19374.

1009 84. Márquez-Jurado, S., Díaz-Colunga, J., das Neves, R.P., Martinez-Lorente, A., Almazán, F., Guantes, R.,
1010 and Iborra, F.J. (2018). Mitochondrial levels determine variability in cell death by modulating apoptotic
1011 gene expression. Nat Commun *9*, 389.

1012 85. Dölle, C., Flønes, I., Nido, G.S., Miletic, H., Osuagwu, N., Kristoffersen, S., Lilleng, P.K., Larsen, J.P.,
1013 Tysnes, O.-B., Haugarvoll, K., et al. (2016). Defective mitochondrial DNA homeostasis in the substantia
1014 nigra in Parkinson disease. Nat Commun *7*, 13548.

1015 86. Chen, C., Turnbull, D.M., and Reeve, A.K. (2019). Mitochondrial Dysfunction in Parkinson's Disease-
1016 Cause or Consequence? Biology (Basel) *8*,.

1017 87. Pinto, M., and Moraes, C.T. (2014). Mitochondrial genome changes and neurodegenerative diseases.
1018 Biochim Biophys Acta *1842*, 1198–1207.

1019 88. Rodríguez-Santiago, B., Casademont, J., and Nunes, V. (2001). Is mitochondrial DNA depletion
1020 involved in Alzheimer's disease? Eur J Hum Genet *9*, 279–285.

1021 89. Chew, H., Solomon, V.A., and Fonteh, A.N. (2020). Involvement of Lipids in Alzheimer's Disease
1022 Pathology and Potential Therapies. Front Physiol *11*, 598.

1023 90. Yin, J., Reiman, E.M., Beach, T.G., Serrano, G.E., Sabbagh, M.N., Nielsen, M., Caselli, R.J., and Shi, J.
1024 (2020). Effect of ApoE isoforms on mitochondria in Alzheimer disease. Neurology *94*, e2404–e2411.

1025 91. El Khoury, J.B., Moore, K.J., Means, T.K., Leung, J., Terada, K., Toft, M., Freeman, M.W., and Luster,
1026 A.D. (2003). CD36 mediates the innate host response to beta-amyloid. J Exp Med *197*, 1657–1666.

1027 92. Lämsä, R., Helisalmi, S., Herukka, S.-K., Tapiola, T., Pirttilä, T., Vepsäläinen, S., Hiltunen, M., and
1028 Soininen, H. (2008). Genetic study evaluating LDLR polymorphisms and Alzheimer's disease. Neurobiol
1029 Aging *29*, 848–855.

1030 93. Tomioka, M., Toda, Y., Mañucat, N.B., Akatsu, H., Fukumoto, M., Kono, N., Arai, H., Kioka, N., and
1031 Ueda, K. (2017). Lysophosphatidylcholine export by human ABCA7. Biochim Biophys Acta Mol Cell Biol
1032 Lipids *1862*, 658–665.

1033 94. Nuytemans, K., Maldonado, L., Ali, A., John-Williams, K., Beecham, G.W., Martin, E., Scott, W.K., and
1034 Vance, J.M. (2016). Overlap between Parkinson disease and Alzheimer disease in ABCA7 functional
1035 variants. Neurol Genet *2*, e44.

1036 95. Crabos, M., Yamakado, T., Heizmann, C.W., Cerletti, N., Bühler, F.R., and Erne, P. (1991). The calcium
1037 binding protein tropomyosin in human platelets and cardiac tissue: elevation in hypertensive cardiac
1038 hypertrophy. Eur J Clin Invest *21*, 472–478.

1039 96. Kostan, J., Salzer, U., Orlova, A., Törö, I., Hodnik, V., Senju, Y., Zou, J., Schreiner, C., Steiner, J.,
1040 Meriläinen, J., et al. (2014). Direct interaction of actin filaments with F-BAR protein pacsin2. EMBO Rep
1041 *15*, 1154–1162.

1042 97. Sever, S. (2002). Dynamin and endocytosis. Curr Opin Cell Biol *14*, 463–467.

1043 98. Cai, B., Giridharan, S.S.P., Zhang, J., Saxena, S., Bahl, K., Schmidt, J.A., Sorgen, P.L., Guo, W.,
1044 Naslavsky, N., and Caplan, S. (2013). Differential roles of C-terminal Eps15 homology domain proteins as
1045 vesiculators and tubulators of recycling endosomes. J Biol Chem *288*, 30172–30180.

1046 99. Di Buduo, C.A., Moccia, F., Battiston, M., De Marco, L., Mazzucato, M., Moratti, R., Tanzi, F., and
1047 Balduini, A. (2014). The importance of calcium in the regulation of megakaryocyte function.
1048 Haematologica *99*, 769–778.

1049 100. De Botton, S., Sabri, S., Daugas, E., Zermati, Y., Guidotti, J.E., Hermine, O., Kroemer, G.,
1050 Vainchenker, W., and Debili, N. (2002). Platelet formation is the consequence of caspase activation
1051 within megakaryocytes. Blood *100*, 1310–1317.

1052 101. Josefsson, E.C., James, C., Henley, K.J., Debrincat, M.A., Rogers, K.L., Dowling, M.R., White, M.J.,
1053 Kruse, E.A., Lane, R.M., Ellis, S., et al. (2011). Megakaryocytes possess a functional intrinsic apoptosis
1054 pathway that must be restrained to survive and produce platelets. J Exp Med *208*, 2017–2031.

1055 102. Josefsson, E.C., Burnett, D.L., Lebois, M., Debrincat, M.A., White, M.J., Henley, K.J., Lane, R.M.,
1056 Moujalled, D., Preston, S.P., O'Reilly, L.A., et al. (2014). Platelet production proceeds independently of
1057 the intrinsic and extrinsic apoptosis pathways. Nat Commun *5*, 3455.

1058 103. McArthur, K., Chappaz, S., and Kile, B.T. (2018). Apoptosis in megakaryocytes and platelets: the life
1059 and death of a lineage. Blood *131*, 605–610.

1060 104. Linsel-Nitschke, P., Götz, A., Erdmann, J., Braenne, I., Braund, P., Hengstenberg, C., Stark, K.,
1061 Fischer, M., Schreiber, S., El Mokhtari, N.E., et al. (2008). Lifelong reduction of LDL-cholesterol related to
1062 a common variant in the LDL-receptor gene decreases the risk of coronary artery disease--a Mendelian
1063 Randomisation study. PLoS One *3*, e2986.

1064 105. Lawlor, D.A., Harbord, R.M., Sterne, J.A.C., Timpson, N., and Davey Smith, G. (2008). Mendelian
1065 randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med *27*,
1066 1133–1163.

1067 106. Sookoian, S., Rosselli, M.S., Gemma, C., Burgueño, A.L., Gianotti, T.F., Castaño, G.O., and Pirola, C.J.
1068 (2010). Epigenetic regulation of insulin resistance in nonalcoholic fatty liver disease: Impact of liver
1069 methylation of the peroxisome proliferator–activated receptor γ coactivator 1α promoter. Hepatology
1070 *52*, 1992–2000.

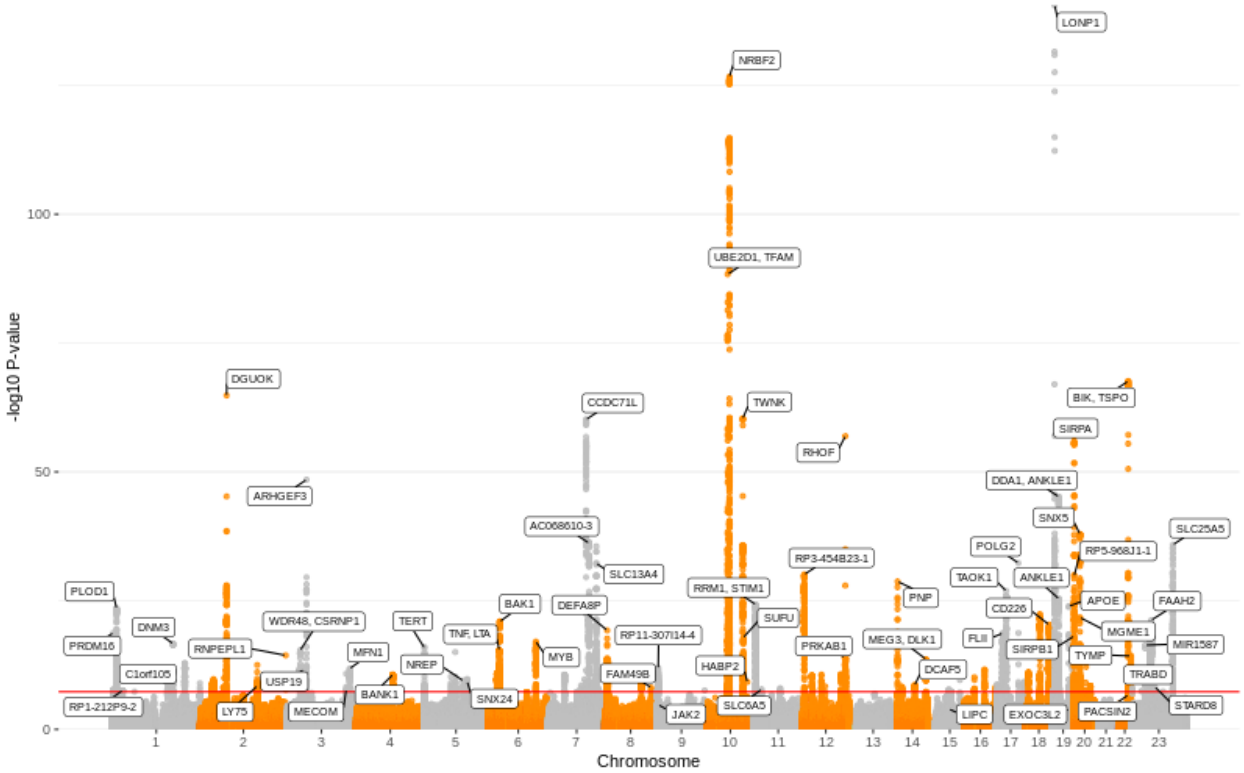
1071 107. Vyas, C.M., Ogata, S., Reynolds, C.F., Mischoulon, D., Chang, G., Cook, N.R., Manson, J.E., Crous-
1072 Bou, M., De Vivo, I., and Okereke, O.I. (2020). Lifestyle and behavioral factors and mitochondrial DNA
1073 copy number in a diverse cohort of mid-life and older adults. PLoS One *15*, e0237235.

1074

1075 **Figures**

1076 **Figure 1. Manhattan plot of GWS loci from UKB-only analyses.**

1077

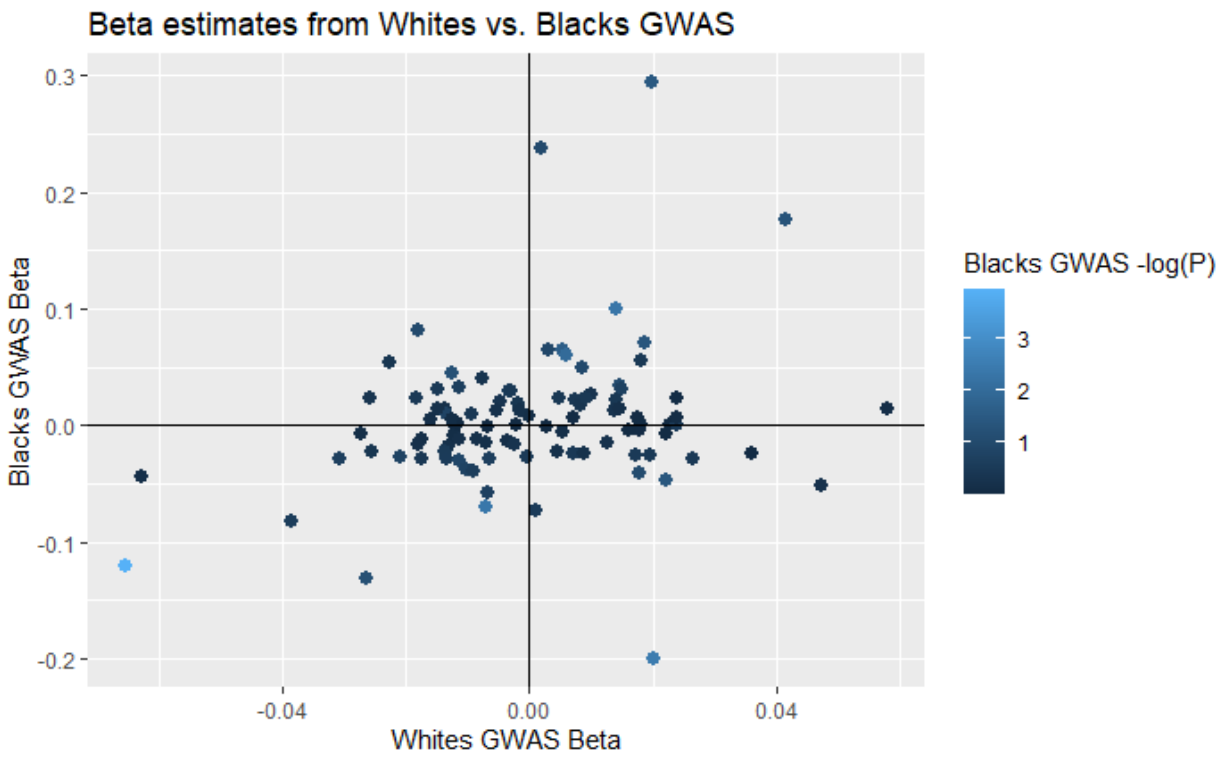


1078

1079 Manhattan plot showing genome-wide significant loci for the UK Biobank-only analyses.

1080

1081 **Figure 2. Scatterplot displaying effect size estimates between Whites and Blacks GWAS results for the**
1082 **129 autosomal SNPs identified in the Whites analyses.**



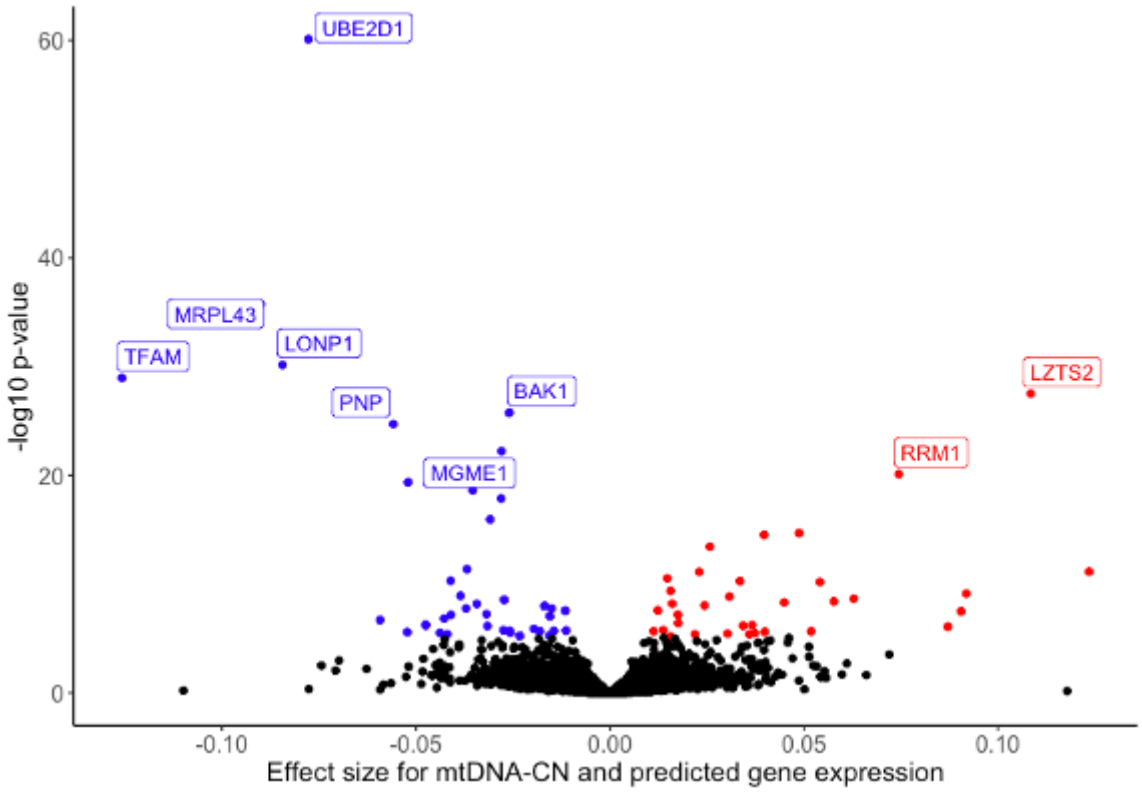
1083

1084 Scatterplot showing comparison between effect size estimates for White and Black individuals. Color
1085 represents significance of effect for each locus in Blacks GWAS analyses.

1086

1087

1088 **Figure 3. Volcano plot of genes whose predicted gene expression is significantly associated with**
1089 **mtDNA-CN.**

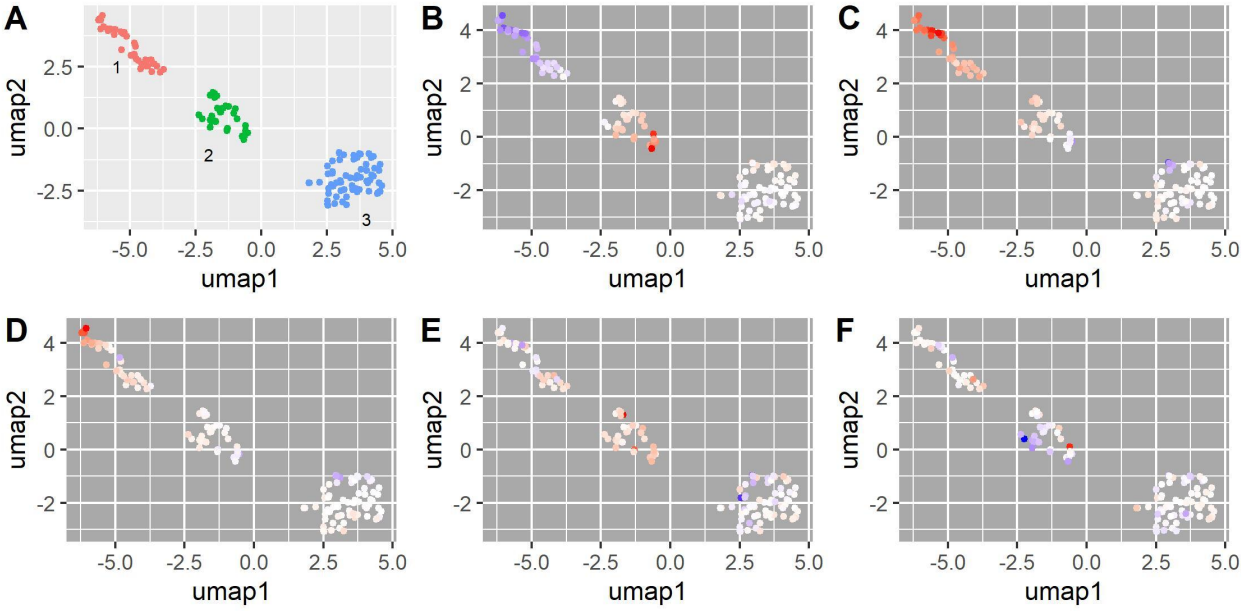


1090

1091 Volcano plot showing genes whose predicted gene expression is significantly associated with mtDNA-CN.
1092 Red indicates positive associations, blue indicates negative associations. Three genes (ARRDC1, EHMT1,
1093 PNPLA7) had extreme effect size estimates greater than 0.3 but were non-significant and removed from
1094 the plot for readability.

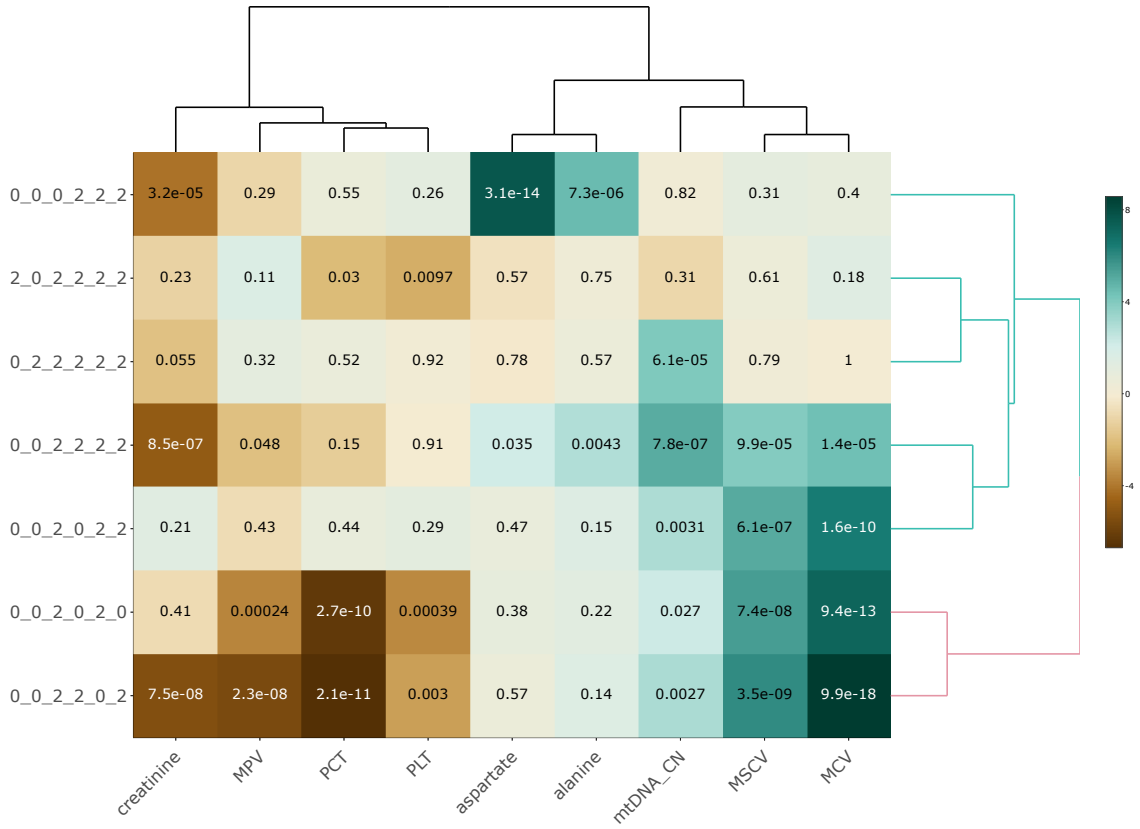
1095

1096 **Figure 4. PHEWAS-based clustering of mtDNA-CN associated SNPs.**



1097
1098
1099 **UMAP clusters created from PHEWAS associations for mtDNA-CN associated SNPs. (A) Three clusters**
1100 **were identified as labeled in the panel; orange indicates no cluster. (B-F) SNPs are colored based on**
1101 **their effect estimate size, standardized to the effect on mtDNA-CN (red = positive, blue = negative**
1102 **estimates), for (B) platelet count, (C) mean platelet volume, (D) platelet distribution width, (E) serum**
1103 **calcium levels, (F) mean spherical cellular volume.**
1104

1105 **Figure 5. Associations between mtDNA-CN associated phenotypes and mitochondrial haplotypes.**



1106

1107 **Mitochondrial haplotypes are significantly associated with mtDNA-CN associated traits, implying**
1108 **causal relationships between mitochondrial function and traits of interest. Haplotypes are notated in**
1109 **the following format: MT73_MT12612_MT7028_MT10238_MT13617_MT15257.**

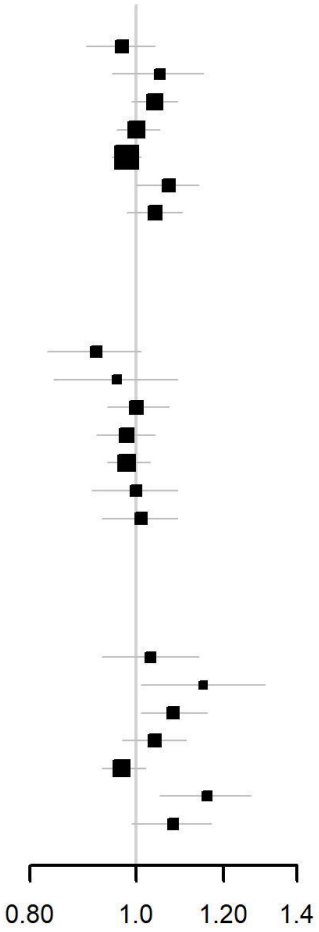
1110

1111

1112 **Figure 6. Associations between mitochondrial DNA haplotypes and morbidity (overall, cancer, and**
1113 **noncancer)**

mtSNP Haplotype Mortality

Haplotype	N	Events	HR (95% CI)
All-cause Mortality			
2_2_2_2_2_2	150,976	9,794	Reference
0_0_0_2_2_2	13,488	855	0.97 (0.90-1.04)
0_0_2_0_2_0	6,899	470	1.05 (0.95-1.15)
0_0_2_0_2_2	34,476	2,318	1.04 (0.99-1.09)
0_0_2_2_0_2	34,843	2,274	1.00 (0.96-1.05)
0_0_2_2_2_2	104,733	6,598	0.98 (0.95-1.01)
0_2_2_2_2_2	14,463	987	1.07 (1.00-1.14)
2_0_2_2_2_2	19,970	1,326	1.04 (0.98-1.10)
Overall P	**379,848**	**24,622**	**0.044**
Cancer Mortality			
2_2_2_2_2_2	150,976	5,324	Reference
0_0_0_2_2_2	13,488	441	0.92 (0.83-1.01)
0_0_2_0_2_0	6,899	233	0.96 (0.84-1.09)
0_0_2_0_2_2	34,476	1,217	1.00 (0.94-1.07)
0_0_2_2_0_2	34,843	1,199	0.98 (0.92-1.04)
0_0_2_2_2_2	104,733	3,614	0.98 (0.94-1.03)
0_2_2_2_2_2	14,463	501	1.00 (0.91-1.09)
2_0_2_2_2_2	19,970	702	1.01 (0.93-1.09)
Overall P	**379,848**	**13,231**	**0.742**
Non-cancer Mortality			
2_2_2_2_2_2	150,976	4,470	Reference
0_0_0_2_2_2	13,488	414	1.03 (0.93-1.14)
0_0_2_0_2_0	6,899	237	1.15 (1.01-1.31)
0_0_2_0_2_2	34,476	1,101	1.08 (1.01-1.16)
0_0_2_2_0_2	34,843	1,075	1.04 (0.97-1.11)
0_0_2_2_2_2	104,733	2,984	0.97 (0.93-1.02)
0_2_2_2_2_2	14,463	486	1.16 (1.05-1.27)
2_0_2_2_2_2	19,970	624	1.08 (0.99-1.17)
Overall P	**379,848**	**11,391**	**0.00066**



1114
1115 **Mitochondrial haplotypes are significantly associated with overall morbidity, in particular, non-cancer**
1116 **mortality. Haplotypes are notated in the following format: MT73_ MT7028_MT10238_**
1117 **MT12612_MT13617_MT15257.**
1118