

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

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

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

<sup>2</sup> Department of Pathology and Laboratory Medicine, Western University, London, ON, Canada

<sup>3</sup> Department of Laboratory Medicine and Pathology, University of Minnesota Medical School, Minneapolis, MN

<sup>4</sup> 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

<sup>5</sup> Cardiovascular Health Research Unit, Departments of Medicine and Biostatistics, University of Washington, Seattle, WA

<sup>6</sup> 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<sup>1,2</sup>. 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<sup>3-5</sup>.

Unlike the nuclear genome, a large amount of variation exists in the number of copies of mtDNA present within cells, tissues, and individuals. The relative copy number of mtDNA (mtDNA-CN) has been shown to be positively correlated with oxidative stress<sup>6</sup>, energy reserves, and mitochondrial membrane potential<sup>7</sup>. As a minimally invasive proxy measure of mitochondrial dysfunction<sup>8</sup>, decreased mtDNA-CN measured in blood has been previously associated with aging-related disease states including frailty<sup>9</sup>, cardiovascular disease<sup>10–12</sup>, chronic kidney disease<sup>13</sup>, neurodegeneration<sup>14,15</sup>, and cancer<sup>16</sup>.

Although mtDNA-CN measured from whole blood presents itself as an easily accessible and minimally invasive biomarker, cell type composition has been shown to be an important confounder, complicating analyses<sup>17,18</sup>. For example, while platelets generally have fewer mtDNA molecules than leukocytes, the lack of a platelet nuclear genome drastically skews mtDNA-CN estimates. As a result, not only is controlling for cell composition extremely vital for accurate mtDNA-CN estimation, interpreting the results in relation to the impact of cell composition becomes a necessity<sup>18–20</sup>.

Although the comprehensive mechanism through which mtDNA-CN is modulated is largely unknown<sup>21,22</sup>, twin studies have estimated a broad-sense heritability of ~0.65, consistent with moderate genetic control<sup>23</sup>. Several nuclear genes have been shown to directly modulate mtDNA-CN, specifically those within the mtDNA replication machinery such as the mitochondrial polymerase, *POLG* and *POLG2*<sup>24,25</sup>, as well as the mitochondrial DNA helicase, *TWINK*, and the mitochondrial single-stranded binding protein, *mtSSB*<sup>26</sup>. Furthermore, nuclear genes which maintain proper mitochondrial nucleotide supply including *DGUOK* and *TK2* have also been shown to regulate mtDNA-CN<sup>27–29</sup>. To further elucidate the genetic control over mtDNA-CN, several genome-wide association studies (GWAS) of mtDNA-CN have been published<sup>30–33</sup>, including a study that was published while the current manuscript was in preparation, analyzing ~300,000 participants from the UK Biobank (UKB), and identifying 50 independent loci<sup>33</sup>.

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<sup>34</sup> and the UK Biobank  
122 (UKB)<sup>35</sup>. 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***

mtDNA-CN was determined using a quantitative PCR assay as previously described<sup>32,36</sup>. Briefly, the cycle threshold (Ct) value of a nuclear-specific and mitochondrial-specific probe were measured in triplicate for each sample. In CHS, a multiplex assay using the mitochondrial *ND1* probe and nuclear *RPPH1* probe was used, whereas ALSPAC used a mitochondrial probe targeting the D-Loop and a nuclear probe targeting *B2M*. In CHS, we observed plate effects, as well as a linear increase in  $\Delta C_t$  due to the pipetting order of each replicate. These effects were corrected in the analysis using linear mixed model regression, with pipetting order included as a fixed effect and plate as a random effect to create a raw measure of mtDNA-CN. In ALSPAC, run-to-run variability was controlled using 3 calibrator samples added to every plate, to allow for adjustment by a per-plate calibration factor<sup>32</sup>.

### **Microarray**

Microarray probe intensities were used to estimate mtDNA-CN using the Genvisis software package<sup>37</sup> as previously described<sup>10,36</sup>. Briefly, Genvisis uses the median mitochondrial probe intensity across all homozygous mitochondrial SNPs as an initial estimate of mtDNA-CN. Technical artifacts such as DNA input quality, DNA input quantity, and hybridization efficiency were captured through either surrogate variable (SV) or principal component (PC) analyses. SVs or PCs were adjusted for through stepwise linear regression by adding successive components until each successive surrogate variable or principal component no longer significantly improved the model.

### **Whole Genome Sequencing (ARIC)**

Whole genome sequencing read counts were used to estimate mtDNA-CN as previously described<sup>36</sup>. Briefly, the total number of reads in a sample were web scraped from the NCBI sequence read archive. Mitochondrial reads were downloaded directly from dbGaP through Samtools (1.3.1). There was no

overlap between ARIC microarray and ARIC whole-genome sequencing samples. A ratio of mitochondrial reads to total aligned reads was used as a raw measure of mtDNA-CN.

### ***Adjusting for Covariates***

Each method described above represents a raw measure of mtDNA-CN, adjusted for technical artifacts; however, several potential confounding variables (e.g., age, sex, blood cell composition) have been identified previously<sup>18</sup>. Raw mtDNA-CN values were adjusted for white blood cell count in ARIC, SHIP and CHS (which also adjusted for platelet count), depending on available data. Standardized residuals (mean = 0, standard deviation = 1) of mtDNA-CN were used after adjusting for covariates (Supplemental Table 1).

### **Estimation of Mitochondrial DNA Copy Number (UKB)**

Due to the availability of more detailed cell count data, as well as a different underlying biochemistry for the Affymetrix Axiom array compared to the genotyping arrays used in the CHARGE cohorts, mtDNA-CN in the UKB was estimated differently (Supplemental Methods). Briefly, mtDNA-CN estimates derived from whole exome sequencing data, available on ~50,000 individuals, were generated first using customized Perl scripts to aggregate the number of mapped sequencing reads and correct for covariates through both linear and spline regression models. Concurrently, mitochondrial probe intensities from the Affymetrix Axiom arrays, available on the full ~500,000 UKB cohort, were adjusted for technical artifacts through principal components generated from nuclear probe intensities. Probe intensities were then regressed onto the whole exome sequencing mtDNA-CN metric, and beta estimates from that regression were used to estimate mtDNA-CN in the full UKB cohort. Finally, we used a 10-fold cross validation method to select the cell counts to include in the final model (Supplemental Table 2). The

final UKB mtDNA-CN metric is the standardized residuals (mean = 0, standard deviation = 1) from a linear model adjusting for covariates (age, sex, cell counts) as described in the Supplemental Methods.

## **Genome-Wide Association Study**

For each individual cohort, regression analysis was performed with residualized mtDNA-CN as the dependent variable adjusting for age, sex, and cohort-specific covariates (e.g., principal components, DNA collection site, family structure, cell composition). Cohorts with multiple mtDNA-CN estimation platforms were stratified into separate analyses. Ancestry-stratified meta-analyses were performed using Metasoft software using the Han and Eskin random effects model to control for unobserved heterogeneity due to differences in mtDNA-CN estimation method<sup>38</sup>. Effect size estimates for SNPs were calculated using a random effect meta-analysis from cohort summary statistics, as the Han and Eskin model relaxes the assumption under the null hypothesis without modifying the effect size estimates that occur under the alternative hypothesis<sup>38</sup>. In total, three complementary analyses were performed in self-identified White individuals: (1) a meta-analysis using all available studies, (2) a meta-analysis of studies with available data for cell count adjustments, and (3) an analysis of UKB-only data. As the vast majority of samples are derived from the UKB study, and given the difficulty in interpreting effect size estimates from a random effects model, further downstream analyses were all performed using effect size estimates from UKB-only data. We additionally performed X chromosome analyses, using only UKB data. X-chromosome analyses were stratified by sex (males = 194,151, females = 216,989), and summary statistics were meta-analyzed using METAL<sup>39</sup> to obtain the final effect estimates.

## **SNP Heritability Estimation**

SNP heritability estimates were retrieved from BOLT-LMM<sup>40</sup>. To verify this metric, we used SumHer<sup>41</sup> to calculate an independent heritability metric using summary statistics. The heritability model used in this

analysis was the BLD-LDAK model. The tagging file used is the pre-computed UK Biobank GBR version for the corresponding heritability model. The summary statistics were filtered so that only single-character reference and alternate alleles are allowed. Chr:BP combination duplicates were removed except the first appearance. SNP heritability was then calculated and extracted from output files.

## **Identification of Independent GWAS Loci**

To identify the initial genome-wide significant (lead) SNPs in each locus, the most significant SNP that passed genome-wide significance ( $p < 5 \times 10^{-8}$ ) within a 1 Mb window was selected. To avoid Type I error, SNPs were only retained for further analyses if there were either (a) at least two genome-wide significant SNPs in the 1 Mb window or (b) if the lead SNP was directly genotyped. Conditional analyses were performed in UKB, where the lead SNPs from the original GWAS were used as additional covariates in order to identify additional independent associations.

## **Comparisons with Hägg et al. 2020**

To compare results with Hägg et al. 2020<sup>33</sup>, summary statistics were obtained from their Supplementary Table 4. Loci were identified as shared between the two GWAS if two lead SNPs were fewer than 500,000 base pairs apart from one another.

## **Fine-mapping**

The susier package was used to identify all potential causal variants for each independent locus associated with mtDNA CN<sup>42</sup>. UKB imputed genotype data for unrelated White subjects were used and variants were extracted using a 500 kb window around the lead SNP for each locus with minor allele frequency (MAF) > 0.001. 95% credible sets (CS) of SNPs, containing a potential causal variant within a locus, were generated. The minimum absolute correlation within each CS is 0.5 and the scaled prior

variance is 0.01. When the CS did not include the lead SNP identified from the GWAS, some of the parameters were slightly relaxed [minimum absolute correlation is 0.2, estimate prior variance is TRUE]. The SNP with the highest posterior inclusion probability (PIP) within each CS was also identified (Supplemental Table 3). With a few exceptions, final lead SNPs were selected by prioritizing initially identified SNPs unless the SNP with the highest PIP had a PIP greater than 0.2 and was 1.75 times larger than the SNP with the second highest PIP.

## **Functional Annotation and Gene Prioritization**

### ***Functional Annotation***

ANNOVAR was used for functional annotation of variants identified in the fine-mapping step<sup>43</sup>. First, variants were converted to an ANNOVAR-ready format using the dbSNP version 150 database<sup>44</sup>. Then, variants were annotated with ANNOVAR using the RefSeq Gene database<sup>45</sup>. The annotation for each variant includes the associated gene and region (e.g., exonic, intronic, intergenic). For intergenic variants, ANNOVAR provides flanking genes and the distance to each gene. For exonic variants, annotations also include likely functional consequences (e.g., synonymous/nonsynonymous, insertion/deletion), the gene affected by the variant, and the amino acid sequence change (Supplemental Table 4).

### ***Co-localization Analyses***

Co-localization analyses were performed using the approximate Bayes factor method in the R package *coloc*<sup>46</sup>. Briefly, *coloc* utilizes eQTL data and GWAS summary statistics to evaluate the probability that gene expression and GWAS data share a single causal SNP (colocalize). *Coloc* returns multiple posterior probabilities; H0 (no causal variant), H1 (causal variant for gene expression only), H2 (causal variant for mtDNA-CN only), H3 (two distinct causal variants), and H4 (shared causal variant for gene expression and mtDNA-CN). In the event of high H4, we designate the gene as causal for the GWAS phenotype of

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

### **DEPICT**

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

### **Gene Assignment**

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

no assigned gene, the most significant DEPICT gene was assigned. If there was no co-localization or DEPICT evidence, the nearest gene was assigned.

## Gene Set Enrichment Analyses

Using the finalized gene list from the prioritization pipeline, GO and KEGG pathway enrichment analyses were performed using the “goana” and “kegga” functions from the R package *limma*<sup>49</sup>, treating all known genes as the background universe<sup>50</sup>. Only one gene per locus was used for “goana” and “kegga” gene set enrichment analysis, prioritizing genes assigned to primary independent hits. If there were multiple assigned genes, one gene was randomly selected to avoid biasing results through loci with multiple functionally related genes. To identify an appropriate p-value cutoff, 100 genes were randomly selected from the genome and run through the same enrichment analysis. This permutation was repeated 1000 times to generate a null distribution of the smallest p-values from each permutation. For cluster-specific gene set enrichment analyses, permutation testing used the same number of random genes as the number of genes in each cluster. To ensure robustness of results, gene set enrichment analysis was repeated 50 times with random selection of genes at loci with multiple assigned genes. GO and KEGG terms that passed permutation cutoffs at least 40/50 times were retained.

## Gene-based Association Test

We used metaXcan, which employs gene expression prediction models to evaluate associations between phenotypes and gene expression<sup>51</sup>. We obtained pre-calculated expression prediction models and SNP covariance matrices, computed using whole blood from European ancestry individuals in version 7 of the Genotype-Tissue expression (GTEx) database<sup>52</sup>. Using prediction performance  $p < 0.05$ , a total of 6,285 genes were predicted. Of these genes, 74 passed Bonferroni correction of  $p < 7.95 \times 10^{-6}$ . Gene set enrichment analyses were performed on Bonferroni-significant genes as previously described.

REVIGO<sup>53</sup> was used on the “medium” setting (allowed similarity = 0.7) to visualize significantly enriched GO terms.

We used a one-sided Fisher’s exact test to test for enrichment of genes that have been previously identified as causal for mtDNA depletion syndromes<sup>54–56</sup>.

## **PHEWAS-based SNP Clustering**

### ***mtDNA-CN Phenome-wide Association Study (PHEWAS)***

We used the PHEnome Scan ANALysis Tool (PHESANT)<sup>57</sup> to identify mtDNA-CN associated quantitative traits in the UKB. Briefly, we tested for the association of mtDNA-CN with 869 quantitative traits (Supplemental Table 5), limiting analyses to 365,781 White, unrelated individuals (used.in.pca.calculation=1). As extreme cell count measurements could indicate individuals with active infections or cancers, they were excluded from analysis (see Supplemental Methods). Analyses were adjusted for age, sex, and assessment center.

### ***SNP-Phenotype Associations***

SNP genotypes were regressed on mtDNA-associated quantitative phenotypic traits using linear regression, adjusted for sex, age with a natural spline (df = 2), assessment center, genotyping array, and 40 genotyping principal components (provided as part of the UKB data download).

### ***SNP Clustering***

To identify distinct clusters of mtDNA-CN GWS SNPs based on phenotypic associations, beta estimates from the SNP-phenotype associations were first divided by the beta estimate of the mtDNA-CN SNP-mtDNA-CN association, so that all SNP-phenotype associations are relative to the mtDNA-CN increasing

allele and scaled to the effect of the SNP on mtDNA-CN. The adjusted beta estimates were subjected to a dimensionality reduction method, Uniform Manifold and Approximation Projection (UMAP), as implemented in the R package *umap*<sup>58</sup> (random\_state=123, n\_neighbors=10, min\_dist=0.001, n\_components=2, n\_epochs=200). SNPs were assigned to clusters using Density Based Clustering of Applications with Noise (DBSCAN) as implemented in the R package *dbscan*<sup>59</sup> (minPts=10). Robustness of cluster assignment was established by varying n\_neighbors, min\_dist, and random\_state parameters. Clusters represent groups of SNPs with similar phenotypic associations.

### ***Phenotype Enrichment and Permutation Testing***

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

### **mtDNA Variant Association Analyses**

#### ***Mitochondrial Variant Phasing and Imputation***

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

panel allele. All genotype files, including the reference panel, were phased using Shapeit4 to fill in any missing genotypes using the phasing iteration sequence “10b,1p,1b,1p,1b,1p,1b,1p,10m”, where b is burn-in iteration, p is pruning iteration, and m is main iteration. The `–sequencing` option was also used due to the presence of multiple mtDNA variants in a very small region, analogous to sequencing data.

Phased UK Biobank genotypes were then imputed with the reference panel using Impute5 with the following parameters: `–pbwt-depth 8`; `–pbwt-cm 0.005`; `–no-threshold`. All imputed variants were functionally annotated using MSeqDR mvTools<sup>63</sup>.

### mtSNP Association Tests

Linear regressions stratified by genotyping array (UKBB, UKBL) were performed for each mtDNA SNP on the 41 traits and mtDNA-CN, including the following covariates: age, age<sup>2</sup>, sex, center, first 20 genotyping PCs. Only SNPs with MAF > 0.005 and imputation INFO score > 0.80 were included (UKBB, n=223; UKBL, n=190; both, n=149). Results were then meta-analyzed using inverse variance weighting. For association analyses between mitochondrial SNPs/haplotypes and mtDNA-CN, the mtDNA-CN metric used was derived only from WES data, as mtDNA genotypes can subtly influence the estimates from genotype array intensities.

### Identification of Independent Genetic Effects

Single SNP study-wide significance was established by generating 300 normally distributed dummy traits, and running single SNP tests using the UKBB data. The minimum SNP p-value for each dummy trait was then selected, and the 15<sup>th</sup> most significant p-value from the 300 analyses was divided by 42 (41 real traits + mtDNA-CN), resulting in study-wide p-value threshold of  $p < 9.5 \times 10^{-6}$ . To identify a subset of traits to perform credible set identification using SusieR (see above, **Fine Mapping**), SNPs were first filtered based on the study-wide p-value threshold, and then then most significantly associated trait

was identified for each SNP. SusieR, (parameters: L=10, estimate\_residual\_variance=TRUE, estimate\_prior\_variance=TRUE, check\_z = FALSE) was then run for each of these traits using the UKBB imputed data and summary association test statistics. A total of 7 credible sets were identified across the 4 traits, two of which co-localized, resulting in 6 credible sets. Independence across the 6 credible sets was tested using multivariate regression models, and requiring  $p < 0.0005$  for at least one trait for a SNP to remain in the model. SNPs MT73A\_G and MT 7028C\_T were in moderate high LD ( $r^2=0.67$ ), but based on conditional regression analyses as described in the main results, capture independent effects and are associated with different traits.

# **Haplotype Generation and Analysis**

Haplotype were constructed by concatenating SNPs across the 6 credible sets using SNPs directly genotyped on both genotyping arrays. This required selecting a SNP with a lower PIP for 2 of the 6 credible sets (MT12612A\_G replaced MT462C\_T,  $r^2 = 0.81$ ; MT10238T\_C replaced MT4529A\_T,  $r^2 = 0.89$ ). Haplotypes with MAF  $< 0.005$  were set to missing ( $n = 1607$ ), resulting in 8 haplotypes, with the most common haplotype set as reference. Significance for haplotype associations with each trait were generated by an anova between regression models with and without the haplotypes. Covariates included age, age<sup>2</sup>, sex, center, first 40 genotyping PCs, and genotyping array.

Mortality analyses were run using Cox proportional hazards models, with covariates as above. Individuals with external causes of death (ICD 10 Death Code categories V, W, X, Y) were censored at time of death. Additionally, for non-cancer mortality analyses, cancer death (ICD 10 Death Code categories C00-D48) were censored. For cancer mortality analyses, all death due to non-cancer cases were censored at time of death.

Clustering for visualization was performed using the R package ‘heatmapply’, with default setting and  
hclust\_method=“ward.D2”.

All statistical analyses were performed using R version 4.0.3.

## Results

### Sample Characteristics

The current study included 465,809 White individuals (53.9% female) with an average age of 56.6 yrs (sd = 8.2 yrs) (Supplemental Table 1). Follow-up validation analyses were performed in 4,770 Blacks (60.2% female) with an average age of 61.2 yrs (sd = 7.4 yrs). The majority of the data originated from the UKB (93%). The bulk of the DNA used for mtDNA-CN estimation was derived from buffy coat (95.5%) while the rest was derived from peripheral leukocytes (2.2%), whole blood (2.3%), or brain (< 0.2%). mtDNA-CN estimated from Affymetrix genotyping arrays consisted of 97.9% of the data while the remainder was derived from qPCR (1.8%) and WGS (0.3%).

### GWAS

Previous work has demonstrated that the method used to measure mtDNA-CN can impact the strength of association<sup>36</sup>. To account for potential differences across studies due to the different mtDNA-CN measurements used, as well as the inclusion of blood cell counts as covariates in only a subset of the cohorts, we took two approaches. First, we used a random effects model to perform meta-analyses, allowing for different genetic effect size estimates across cohorts. Second, we performed three complementary analyses in individuals who self-identified as White: 1) meta-analysis of all available studies (n = 465,809); 2) meta-analysis of studies with available data for cell count adjustment (n = 456,151); and 3) GWAS of UKB only (n = 440,266) (Figure 1). 77 loci were significant in

all three meta-analyses, and we identified 93 independent loci that were significant in at least one of the analyses. In the meta-analysis of UKB-only data, 92 of the total 93 loci were identified (Supplemental Figure 3). Given that >90% of the samples come from the UKB study, and the challenge of interpreting effect size estimates from a random effects model, downstream analyses all use effect size estimates from the UKB only analyses (Supplemental Table 6), which showed no evidence for population substructure inflating test statistics, with a genomic inflation factor of 1.09 (Supplemental Figure 4). SNP heritability estimated from BOLT-LMM<sup>40</sup> for mtDNA-CN adjusted for age and sex was 10.5% while heritability for mtDNA-CN adjusted for age, sex, and cell counts was 7.4%, implying that some of the mtDNA-CN heritability observed in previous studies could be due to heritability of cell type composition. We also used SumHer<sup>41</sup> as an alternative approach to calculating SNP heritability, which returned a comparable estimate of 7.0% for the cell-count corrected mtDNA-CN metric.

The most significant SNP associated with mtDNA-CN was a missense mutation in *LONP1* ( $p = 3.00 \times 10^{-141}$ ), a gene that encodes a mitochondrial protease that can directly bind mtDNA, and has been shown to regulate *TFAM*, a transcription factor involved in mtDNA replication and repair (for review see Gibellini *et al.*)<sup>64</sup>.

Meta-analysis of the sex-stratified X chromosome results identified four loci significantly associated with mtDNA-CN, with directionality consistent across the male and female stratified analyses (Supplemental Table 7).

### **Fine-mapping and Secondary Hits**

To identify additional independent SNPs within novel loci whose effects were masked by the original significant SNP, as well as identify additional loci, we took two approaches. First, a conditional analysis adjusting for the top 93 SNPs from the initial (primary) GWAS run revealed 3 novel loci and 19 additional independent significant SNPs within existing loci. We also performed fine-mapping with susieR<sup>42</sup> and

discovered an additional 14 independent SNPs within existing loci. The majority of loci had only one 95% credible set of SNPs; further, twenty of the credible sets contained only one SNP. However, many of the credible sets contained greater than 50 SNPs after fine-mapping, and 12 of the 122 credible sets had a missense SNP as the SNP with the highest PIP in the set. Using these two methods, we identified in total 129 independent SNPs across 96 autosomal loci (Supplemental Figure 5), while susieR fine-mapping and conditional analyses for the X-chromosome loci did not reveal any additional secondary signals.

### **Comparisons with Hägg et al. 2020**

Out of the 50 loci reported in Hägg et al. 2020<sup>33</sup>, we replicate 38 loci in our cell-count adjusted analyses (Supplemental Table 8). As the two GWASs both use UK Biobank data, this replication is unsurprising. Out of the 12 loci that were not genome-wide significant in our cell-count adjusted analyses, 11 were significant when we did not adjust our mtDNA-CN metric for cell counts, suggesting that cell-type composition may be driving these signals. The current manuscript also reports 62 additional loci that are not in the Hägg et al. 2020 study. This is likely due to increased power, as the sample size used for the current analyses is nearly twice as large.

### **Associations in Black Populations**

Examining the 129 autosomal SNPs from the Whites-only analysis, 99 were available in the Blacks-only meta-analysis (n = 4770). After multiple testing correction, one of these SNPs was significant (rs73349121,  $p = 0.0001$ ), 9 were nominally significant ( $p < 0.05$ , with 5 expected), and 58/99 had a direction of effect that was consistent with the Whites-only analyses (one-sided  $p = 0.04$ , Figure 2). Despite being under-powered, these results in the Blacks-only analyses provide evidence for similar genetic effects in a different ancestry group.

## Gene Prioritization and Enrichment of mtDNA Depletion Syndrome Genes

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

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

## Gene Set Enrichment Analyses

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

replication, and amyloid-beta clearance (Supplemental Table 13). No KEGG terms were significant across multiple iterations.

## MetaXcan Gene Expression Analysis

As a complementary approach to single-SNP analyses, we explored the associations between mtDNA-CN and predicted gene expression using MetaXcan<sup>51</sup>. MetaXcan incorporates multiple SNPs within a locus along with a reference eQTL dataset to generate predicted gene expression levels. As our study estimated mtDNA-CN derived from blood, we used whole blood gene expression eQTLs from the Gene-Tissue Expression (GTEx) consortium<sup>65</sup> to predict gene expression in the UKB dataset. We identified 6,285 genes that had a predicted performance  $p < 0.05$  (i.e., they had sufficient data to generate robust gene expression levels) and were tested for association with mtDNA-CN. Of these genes, 74 were significantly associated with mtDNA-CN ( $p < 7.95 \times 10^{-6}$ ) (Figure 3, Supplemental Table 14), including 8 that were not identified through single-SNP analyses. Many of the significant genes have known mitochondrial functions, notably the mtDNA transcription factor *TFAM* ( $p = 1.09 \times 10^{-29}$ ) and mitochondrial exonuclease *MGME1* ( $p = 5.87 \times 10^{-23}$ ), genes known as causal for mtDNA depletion syndromes<sup>54,55</sup>. Additionally, *LONP1*, *MRPL43*, and *BAK1*, are all genes with known mitochondrial functions<sup>66–68</sup>. Bonferroni significant MetaXcan genes were used for gene enrichment analysis, finding enrichment for “nucleobase metabolic process” ( $p = 1.47 \times 10^{-4}$ ) and “mitochondrial fusion” ( $p = 1.86 \times 10^{-4}$ ) (Supplemental Figure 6).

## PHEWAS-based SNP Clustering and Gene Set Enrichment

mtDNA-CN is associated with numerous quantitative and qualitative phenotypes, many of which are relevant to aging-related disease<sup>3–5,9,10,13–16</sup>. We hypothesized that this pleiotropy may reflect different underlying functional domains captured by mtDNA-CN, and may be reflected in GWAS-identified SNPs

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

#### **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 ( $p < 9.5 \times 10^{-6}$ ), 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<sup>9</sup>. 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<sup>16,72,73</sup>, 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.

We note that during the preparation of this manuscript, a GWAS for mtDNA-CN performed in 295,150 unrelated individuals from the UK Biobank was published, which reported 50 genome-wide significant regions<sup>33</sup>. Within our GWAS, we replicate 38 of these 50 genome-wide significant loci in our cell count corrected analyses. An additional 11 out of the remaining 12 loci are genome-wide significant when we do not adjust mtDNA-CN for cell count. While Hagg et. al adjust for cell type composition, this difference suggests that their adjustment may not be fully capturing the effects of cell counts. Additionally, our analyses report 59 additional loci that are not observed in the previous paper, largely due to the increased power of our study.

We were able to identify a substantial proportion of the genes involved in mtDNA depletion syndromes (7/16,  $p = 3.09 \times 10^{-15}$  for enrichment), including *TWINK*, *TFAM*, *DGUOK*, *MGME1*, *RRM2B*, *TYMP*, and *POLG*. mtDNA depletion syndromes can be broken down into 5 subtypes based on their constellation of phenotypes<sup>74</sup>, and with the exception of cardiomyopathic subtypes (associated with mutations in *AGK* and *SLC25A4*), we were able to identify at least 1 gene from the other 4 subtypes, suggesting that our mtDNA-CN measurement in blood-derived DNA can identify genes widely relevant to non-blood phenotypes. This finding is consistent with a large body of work showing that mtDNA-CN measured in blood is associated with numerous aging-related phenotypes for which the primary tissue of interest is not blood (e.g. chronic kidney disease<sup>13</sup>, heart failure<sup>11</sup>, and diabetes<sup>75</sup>). Also consistent with this finding is recent work demonstrating that mtDNA-CN measured in blood is associated with mtRNA expression across numerous non-blood tissues, suggesting a link between mitochondrial function measured in blood and other tissues<sup>76</sup>.

In addition to identifying the mtDNA depletion syndrome genes directly linked to mitochondrial DNA metabolic processes, DNA replication, and genome maintenance, we also identify genes which play a role in mitochondrial function. The top GWAS hit is a missense mutation in *LONP1*, which encodes a mitochondrial protease that has been shown to cause mitochondrial cytopathy and reduced respiratory

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

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

Given the integral role of mitochondria in cellular function, from ATP formation and energy production, signaling through reactive oxygen species, and apoptosis mediation, there is a strong basis to *a priori* assume that genetic variants associated with mtDNA-CN are likely to be highly pleiotropic. Indeed, mtDNA-CN itself is associated with numerous phenotypes (Supplemental Table 5). Through our PHEWAS-based clustering approach using 41 mtDNA-CN associated phenotypes, we uncovered phenotypic associations between three distinct clusters of GWS mtDNA-CN associated SNPs. Cluster 1 was characterized by increased MPV, PDW, and decreased PLT (note that measured MPV and PLT are generally inversely correlated to maintain hemostasis), which are the hallmarks of platelet activation<sup>69</sup>. The link between platelets and mtDNA-CN has typically revolved around platelet count, as platelets have functional mitochondria, but do not have a nucleus. Given that the mtDNA-CN measurement is the ratio between mtDNA and nuclear DNA, increased platelets, all else being equal, would directly equate with increased mtDNA-CN. We note that the mtDNA-CN metric used in this GWAS was adjusted for platelet count, likely increasing the ability to detect variants that impact mtDNA-CN through increased platelet activation. Examining the genes within this cluster suggests roles for actin formation and regulation (*TPM4*, *PACSIN2*)<sup>95,96</sup> and vesicular transport and endocytic trafficking (*DNM3*, *EHD3*)<sup>97,98</sup> in platelet activation.

Cluster 2 is most strongly enriched for SNPs in which the mtDNA-CN increasing allele is associated with increased PLT/PCT and serum calcium/phosphate. Examining the genes assigned to the cluster, we implicate megakaryocyte proliferation and proplatelet formation (*MYB*, *JAK2*)<sup>70</sup>, and apoptosis and autophagy (*BAK1*, *BCL2*, *TYMP*)<sup>71</sup>. Megakaryocytes are used to form proplatelets, and the process includes an important role for both intra- and extracellular calcium levels<sup>99</sup>. A role for apoptosis, and specifically *BCL2*, in proplatelet formation and platelet release has been suggested<sup>100,101</sup>, however work in mice has suggested that apoptosis does not play a direct role in these processes<sup>102</sup>. Nevertheless, apoptosis is important for platelet lifespan<sup>103</sup>.

Cluster 3 was particularly challenging to interpret, given that no particular phenotype was enriched relative to the non-cluster 3 SNPs. We note that this cluster appeared to be enriched for the mtDNA depletion syndrome genes, containing 6/7 genes identified in the GWAS, and significantly enriched for GO Terms related mitochondrial DNA. Additionally, genes in cluster 3 were significantly enriched for low-density lipoprotein particle binding, suggesting a role for lipid homeostasis. Closer inspection of cluster 3 genes reveals a number of genes known to be associated with lipid levels (*LIPC*, *CETP*, *LDLR*, *APOE*). While lipids play a role in both energy metabolism (largely through fatty acids) and cellular membrane formation, a link to mtDNA-CN and/or mitochondrial function is not well-established.

A strong rationale for the study of mtDNA-CN is the underlying assumption that it reflects mitochondrial function and is readily measured, often from existing data. A serious complication to the interpretation of the role of mitochondrial function in various traits has been use of blood-derived measurements, which can be confounded by differences in cell counts across individuals. Mendelian randomization has been widely used to infer causality between traits (e.g. LDL and CAD)<sup>104</sup>, but is only robust under conditions of little to no pleiotropy<sup>105</sup> and its power is a function of variance explained. For mtDNA-CN, the extensive pleiotropy and small amount of variance explained of GWS variants (<1%) prevents the use of traditional MR approaches. As an alternative approach, we analyzed associations between mitochondrial DNA variants and mtDNA-CN associated phenotypes. Presumably, variants located on the mitochondrial genome are only able to modify phenotypes through modulating mitochondrial function, allowing for causal inference. Our analyses revealed significant relationships between mitochondrial variants and creatinine, aspartate aminotransferase, MCV, and PCT. Creatinine and aspartate aminotransferase are markers of kidney and liver function respectively, and supporting these findings, mtDNA-CN has been linked to both chronic kidney disease<sup>13</sup> and non-alcoholic fatty liver disease<sup>106</sup>. We also find a highly significant association between mitochondrial variation and non-cancer

mortality, adding evidence for a causal relationship to previous findings showing mtDNA-CN is associated with all-cause mortality<sup>9</sup>.

Several limitations should be noted. First, despite the large sample size and numerous loci identified, we are likely missing a great deal of the true signal, as our SNP heritability estimates through SumHer and BOLT-LMM were 7.0% and 7.4% respectively, while previous studies have estimated mtDNA-CN heritability to be 65%<sup>23</sup>. Finally, while we have adjusted our mtDNA-CN metric for a variety of confounders, it is important to note that mtDNA-CN can be influenced by a variety of environmental factors including smoking<sup>107</sup> and drugs, which have not been adjusted for in these analyses.

In summary, we performed the largest-to-date GWAS for mtDNA-CN, including almost 500,000 individuals. We identified three distinct groups of SNPs associated with mtDNA-CN that are related to platelet activation, megakaryocyte formation and apoptotic processes, and showed clear enrichment for genes involved in mtDNA depletion and nucleotide regulation. Additionally, we find that mitochondrial variants are significantly associated with creatinine, aspartate aminotransferase, MCV, and PCT, implying a causal relationship between mitochondrial function and these phenotypes. Finally, we provide strong evidence that mitochondrial function is causal for non-cancer mortality. Given the role of mtDNA-CN, and, by proxy, mitochondrial function in aging-related disease, this work begins to unravel the many varied underlying mechanisms through which mitochondrial function impacts human health.

## **Supplemental Data**

Supplemental Data include six supplemental figures, twenty-two supplemental tables, and supplemental methods.

## **Declaration of Interests**

Psaty serves on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. All other authors declare no competing interests.

## Acknowledgements

This work was supported by National Heart, Lung and Blood Institute, National Institutes of Health (NIH) grants R01HL13573 and R01HL144569 (RJL, SY, CAC, DEA), NIH grant P01-AG027734 (GA, YK, NB, AB) and the National Center for Advancing Translational Sciences, NIH, through Grant KL2TR002490 (LMR). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. LMR was also funded by T32 HL129982.

This research was conducted using data from the Genotype-Tissue Expression (GTEx) project (dbGaP accession: phs000424.v8.p2). The GTEx project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

This research was also conducted using the UK Biobank Resource under Application Number 17731.

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

The FHS phenotype-genotype analyses were supported by the National Institute of Aging (U34AG051418). This research was conducted in part using data and resources from the Framingham Heart Study of the National Heart Lung and Blood Institute of the National Institutes of Health and Boston University School of Medicine. This work was partially supported by the National Heart, Lung and Blood Institute's Framingham Heart Study (Contract No. N01-HC-25195, HHSN268201500001) and its contract with Affymetrix, Inc for genotyping services (Contract No. N02-HL-6-4278). Genotyping, quality control and calling of the Illumina HumanExome BeadChip in the Framingham Heart Study was supported by funding from the National Heart, Lung and Blood Institute Division of Intramural Research (Daniel Levy and Christopher J. O'Donnell, Principle Investigators). The authors thank the participants for their dedication to the study. The authors are pleased to acknowledge that the computational work reported on in this paper was performed on the Shared Computing Cluster which is administered by Boston University's Research Computing Services. URL: [www.bu.edu/tech/support/research/](http://www.bu.edu/tech/support/research/).

MESA and the MESA SHARe projects (dbGaP accession: phs000209.v13.p3) are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278. Genotyping was performed at Affymetrix (Santa Clara, California, USA) and the Broad Institute of Harvard and MIT (Boston, Massachusetts, USA) using the Affymetrix Genome-Wide Human SNP Array 6.0. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive

and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.

ROS/MAP is supported by the Translational Genomics Research Institute and National Institute on Aging (NIA) through grants U01AG46152, U01AG61256, P30AG10161, R01AG17917, RF1AG15819, R01AG30146.

SHIP and SHIP-TREND are part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grants no. 01ZZ9603, 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania, and the network ‘Greifswald Approach to Individualized Medicine (GANI\_MED)’ funded by the Federal Ministry of Education and Research (grant 03IS2061A).

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

## Web Resources

REVIGO was accessed at <http://revigo.irb.hr/>.

## Data and Code Availability

All data used in this manuscript is available through either the UKBiobank and CHARGE consortiums.

Code and scripts are available in a zipped file at <https://www.arkinglab.org/resources/>.

# References

1. Wallace, D.C. (1992). Diseases of the Mitochondrial Dna. *Annual Review of Biochemistry* *61*, 1175–1212.
2. Vakifahmetoglu-Norberg, H., Ouchida, A.T., and Norberg, E. (2017). The role of mitochondria in metabolism and cell death. *Biochem. Biophys. Res. Commun.* *482*, 426–431.
3. Dai, D.-F., Rabinovitch, P.S., and Ungvari, Z. (2012). Mitochondria and cardiovascular aging. *Circ. Res.* *110*, 1109–1124.
4. Cui, H., Kong, Y., and Zhang, H. (2012). Oxidative stress, mitochondrial dysfunction, and aging. *J Signal Transduct* *2012*, 646354.
5. Herst, P.M., Rowe, M.R., Carson, G.M., and Berridge, M.V. (2017). Functional Mitochondria in Health and Disease. *Front Endocrinol (Lausanne)* *8*, 296.
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-related alteration of the copy number of mitochondrial DNA in human leukocytes. *Free Radic Res* *37*, 1307–1317.
7. Guha, M., and Avadhani, N.G. (2013). Mitochondrial retrograde signaling at the crossroads of tumor bioenergetics, genetics and epigenetics. *Mitochondrion* *13*, 577–591.
8. Malik, A.N., and Czajka, A. (2013). Is mitochondrial DNA content a potential biomarker of mitochondrial dysfunction? *Mitochondrion* *13*, 481–492.
9. Ashar, F.N., Moes, A., Moore, A.Z., Grove, M.L., Chaves, P.H.M., Coresh, J., Newman, A.B., Matteini, A.M., Bandeen-Roche, K., Boerwinkle, E., et al. (2015). Association of mitochondrial DNA levels with frailty and all-cause mortality. *J. Mol. Med.* *93*, 177–186.
10. Ashar, F.N., Zhang, Y., Longchamps, R.J., Lane, J., Moes, A., Grove, M.L., Mychaleckyj, J.C., Taylor, K.D., Coresh, J., Rotter, J.I., et al. (2017). Association of Mitochondrial DNA Copy Number With Cardiovascular Disease. *JAMA Cardiol* *2*, 1247–1255.
11. Hong, Y.S., Longchamps, R.J., Zhao, D., Castellani, C.A., Loehr, L.R., Chang, P.P., Matsushita, K., Grove, M.L., Boerwinkle, E., Arking, D.E., et al. (2020). Mitochondrial DNA Copy Number and Incident Heart Failure: The Atherosclerosis Risk in Communities (ARIC) Study. *Circulation* *141*, 1823–1825.
12. Zhao, D., Bartz, T.M., Sotoodehnia, N., Post, W.S., Heckbert, S.R., Alonso, A., Longchamps, R.J., Castellani, C.A., Hong, Y.S., Rotter, J.I., et al. (2020). Mitochondrial DNA copy number and incident atrial fibrillation. *BMC Medicine* *18*, 246.
13. Tin, A., Grams, M.E., Ashar, F.N., Lane, J.A., Rosenberg, A.Z., Grove, M.L., Boerwinkle, E., Selvin, E., Coresh, J., Pankratz, N., et al. (2016). Association between Mitochondrial DNA Copy Number in Peripheral Blood and Incident CKD in the Atherosclerosis Risk in Communities Study. *J Am Soc Nephrol* *27*, 2467–2473.

832 14. Pyle, A., Anugraha, 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. Vösa, 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., Spemulli, 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.

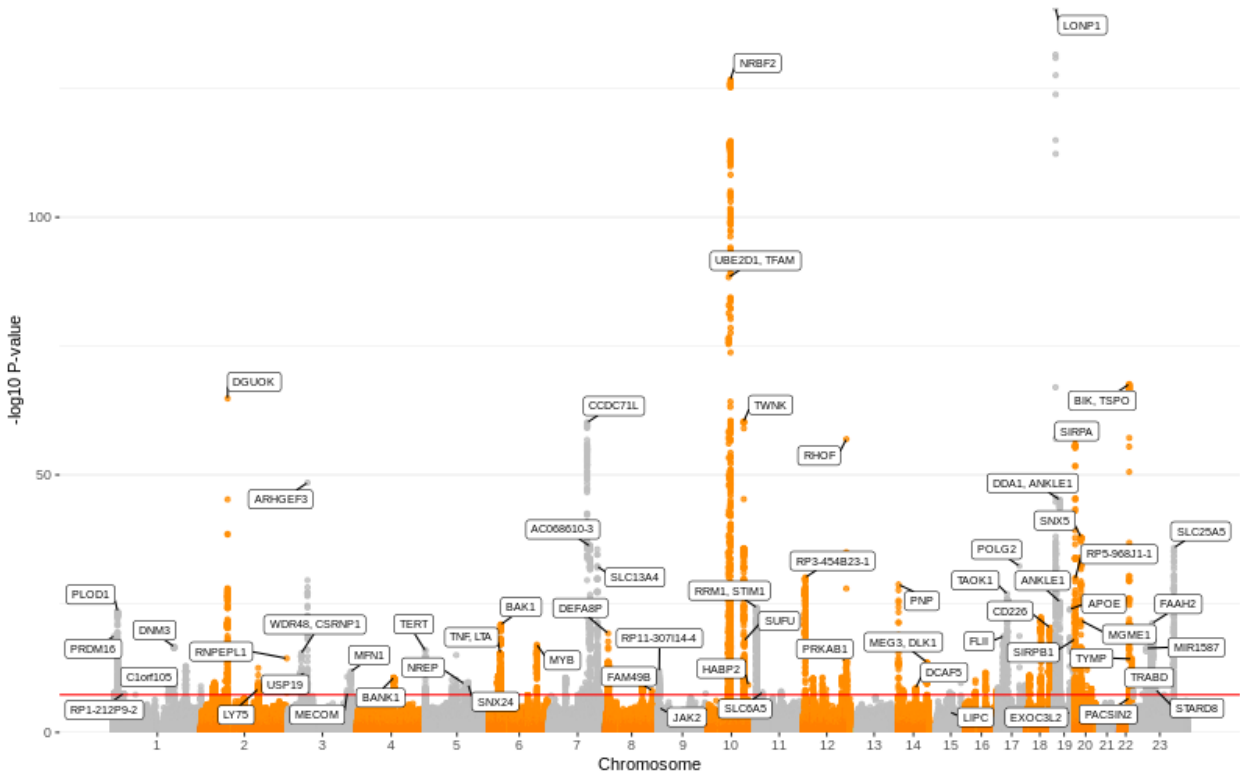
1064 105. Lawlor, D.A., Harbord, R.M., Sterne, J.A.C., Timpson, N., and Davey Smith, G. (2008). Mendelian  
1065 randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med* *27*,  
1066 1133–1163.

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  $\gamma$  coactivator 1 $\alpha$  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.

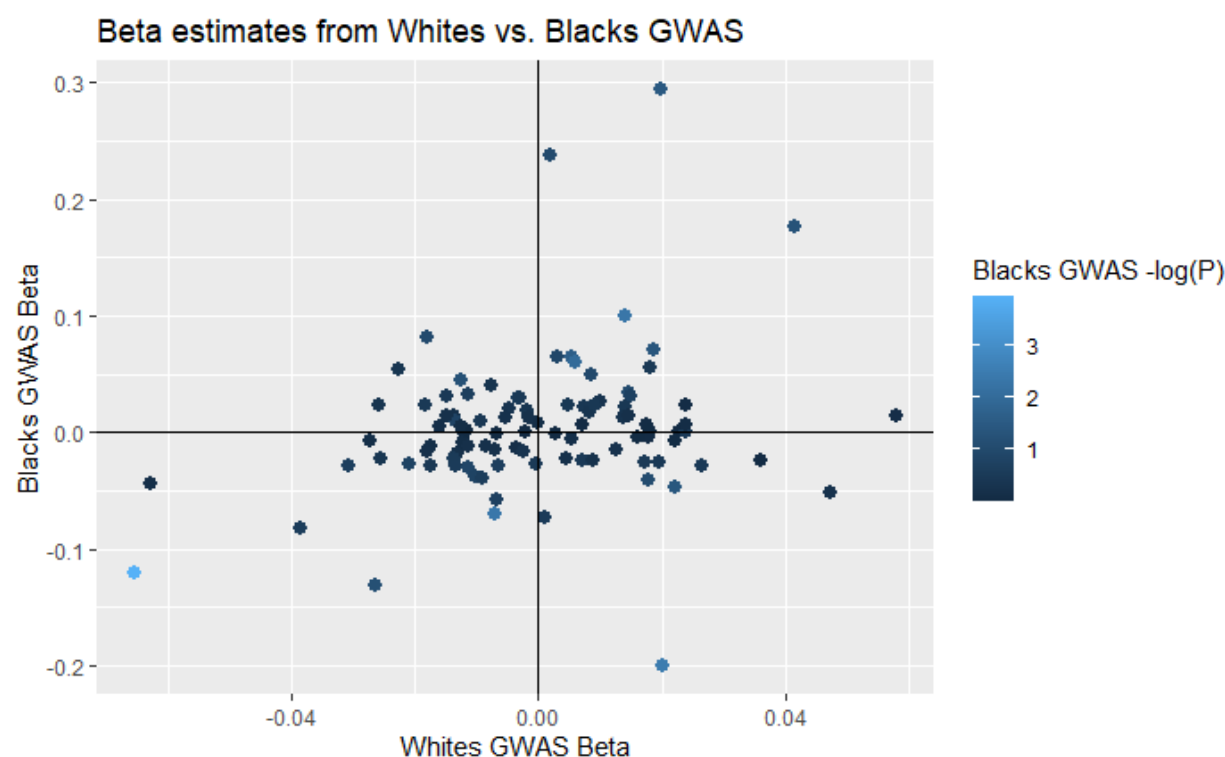
Figures

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



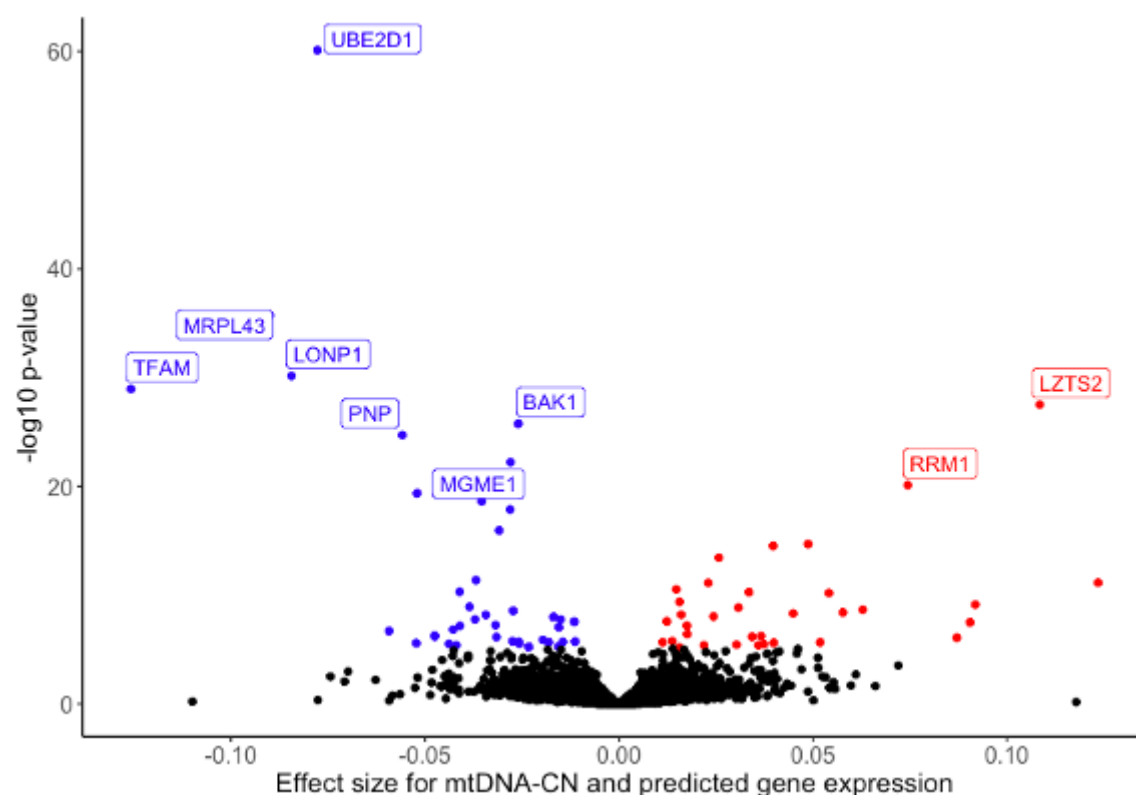
Manhattan plot showing genome-wide significant loci for the UK Biobank-only analyses.

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



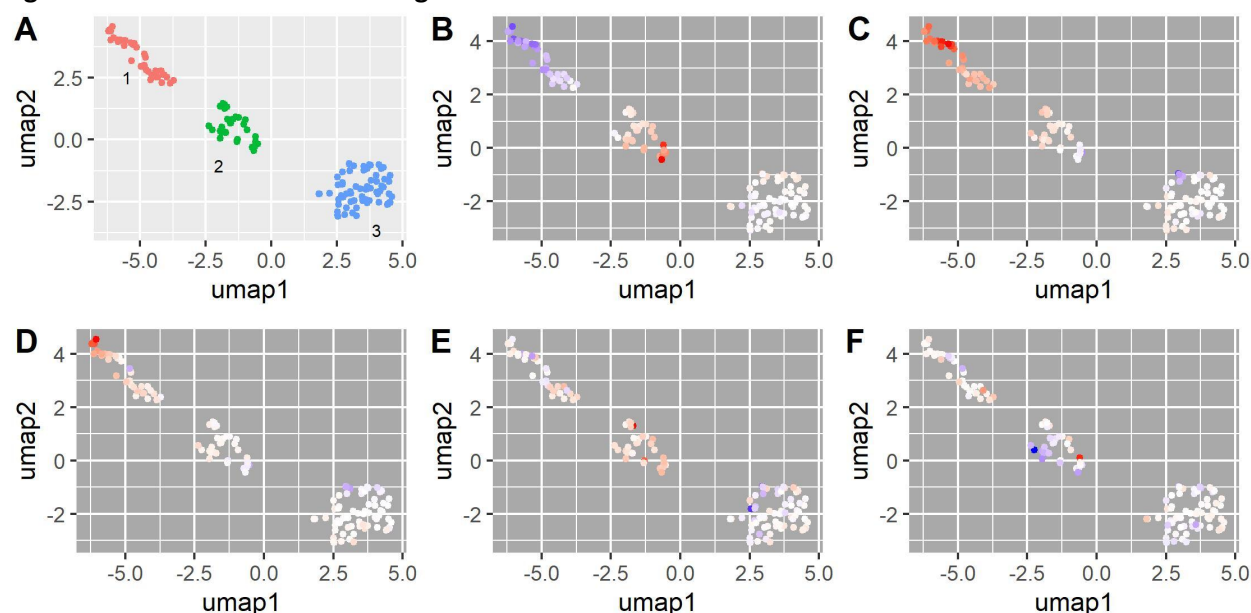
Scatterplot showing comparison between effect size estimates for White and Black individuals. Color represents significance of effect for each locus in Blacks GWAS analyses.

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



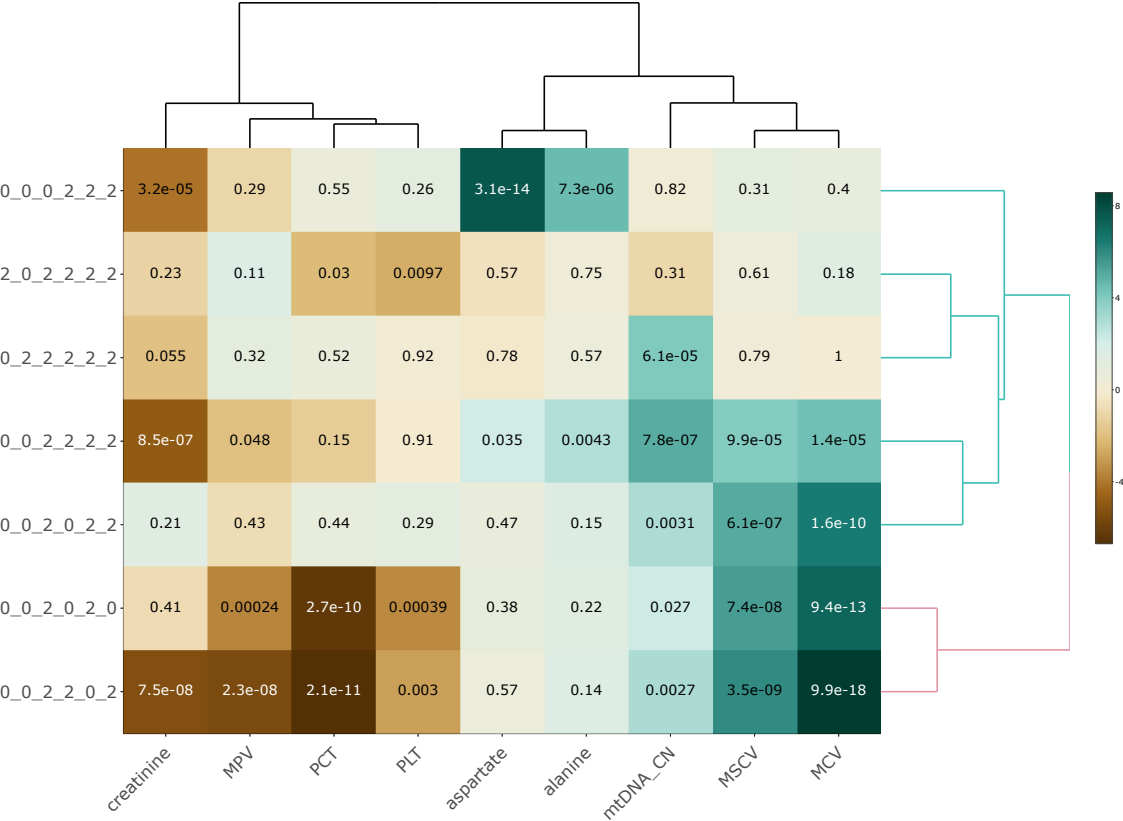
Volcano plot showing genes whose predicted gene expression is significantly associated with mtDNA-CN. Red indicates positive associations, blue indicates negative associations. Three genes (ARRDC1, EHMT1, PNPLA7) had extreme effect size estimates greater than 0.3 but were non-significant and removed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

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



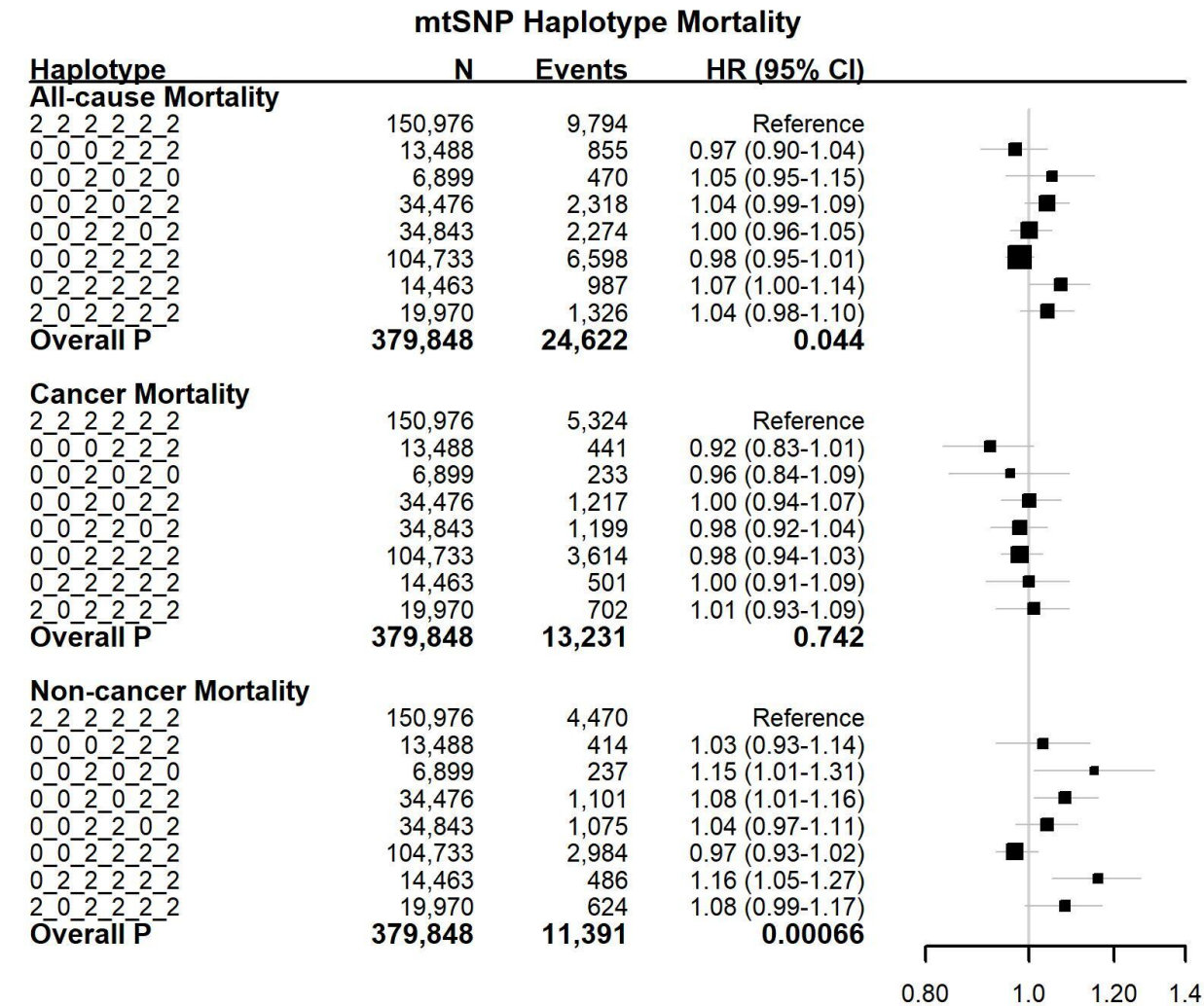
**UMAP clusters created from PHEWAS associations for mtDNA-CN associated SNPs. (A) Three clusters were identified as labeled in the panel; orange indicates no cluster. (B-F) SNPs are colored based on their effect estimate size, standardized to the effect on mtDNA-CN (red = positive, blue = negative estimates),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

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



**Mitochondrial haplotypes are significantly associated with mtDNA-CN associated traits, implying causal relationships between mitochondrial function and traits of interest. Haplotypes are noted in the following format: MT73\_MT12612\_MT7028\_MT10238\_MT13617\_MT15257.**

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



Mitochondrial haplotypes are significantly associated with overall morbidity, in particular, non-cancer mortality. Haplotypes are notated in the following format: MT73\_ MT7028\_ MT10238\_ MT12612\_ MT13617\_ MT15257.