Whole-exome sequencing in 415,422 individuals identifies rare variants associated with mitochondrial DNA copy number

Vamsee Pillalamarri,1,2,3 Wen Shi,1 Conrad Say,1 Stephanie Yang,1,4 John Lane,5 Eliseo Guallar,6 Nathan Pankratz,5 and Dan E. Arking1,7,*

Summary

Inter-individual variation in the number of copies of the mitochondrial genome, called mitochondrial DNA copy number (mtDNA-CN), reflects mitochondrial function and has been associated with various aging-related diseases. We examined 415,422 exomes of self-reported White ancestry individuals from the UK Biobank and tested the impact of rare variants, at the level of single variants and through aggregate variant-set tests, on mtDNA-CN. A survey across nine variant sets tested enrichment of putatively causal variants and identified 14 genes at experiment-wide significance and three genes at marginal significance. These included associations at known mtDNA depletion syndrome genes (mtDNA helicase TWNK, $p = 1.1 \times 10^{-20}$; mitochondrial transcription factor TEAM, $p = 4.3 \times 10^{-15}$; mtDNA maintenance exonuclease MGME1, $p = 2.0 \times 10^{-6}$) and the V617F dominant gain-of-function mutation in the tyrosine kinase JAK2 ($p = 2.7 \times 10^{-11}$), associated with myeloproliferative disease. Novel genes included the ATP-dependent protease CLPX ($p = 8.4 \times 10^{-8}$), involved in mitochondrial proteome quality, and the mitochondrial adenylate kinase AK2 ($p = 4.7 \times 10^{-8}$), involved in hematoopoiesis. The most significant association was a missense variant in SAMHD1 ($p = 4.2 \times 10^{-28}$), found on a rare, 1.2-Mb shared ancestral haplotype on chromosome 20. SAMHD1 encodes a cytoplasmic host restriction factor involved in viral defense response and the mitochondrial nucleotide salvage pathway, and is associated with Aicardi-Goutières syndrome 5, a childhood encephalopathy and chronic inflammatory response disorder. Rare variants were enriched in Mendelian mtDNA depletion syndrome loci, and these variants implicated core processes in mtDNA replication, nucleoid structure formation, and maintenance. These data indicate that strong-effect mutations from the nuclear genome contribute to the genetic architecture of mtDNA-CN.

Introduction

Mitochondria play vital roles in meeting the chemical energy requirements for the cell through oxidative phosphorylation (OXPHOS), in intrinsic apoptosis activation, and in regulation of innate immunity. Most mitochondrial form and function, including factors required to construct the organelle, is encoded in the nuclear genome, but mitochondria have evolved to retain a haploid circular, double-stranded, intron-free, ~16.6-kb DNA molecule (mtDNA) that encodes 13 proteins required for cellular respiration, including most of the components of OXPHOS, and 24 RNAs required for mitochondrial translational machinery. Thus, defects in human mitochondrial genomes, at levels considered heteroplasmic and homoplasmic, contribute to disease risk, including Mendelian mtDNA depletion syndromes and aging-related disease.

Unlike the fixed content of nuclear DNA, mtDNA ploidy is variable in cells and across tissues, individuals, and cohorts. A portion of inter-individual variation in mitochondrial DNA copy number (mtDNA-CN) reflects mitochondrial function and is heritable under nuclear genetic control via common polygenic variation. As a minimally invasive biomarker, mtDNA-CN is associated with several aging-relevant traits, including cardiovascular disease, frailty and all-cause mortality, neurodegeneration, and cancer.

Genome-wide association studies (GWAs) for mtDNA-CN have revealed mtDNA-CN to be a complex phenotype with SNP-based heritability estimates between 7% and 8%. Family-based studies, however, have estimated broad sense heritability for mtDNA-CN to be 54%–65% and variance explained by common independent variants in a non-discovery cohort was 1%–2%. Rare genetic variants, especially mutations in protein-coding regions of the nuclear genome, collectively account for a substantial proportion of missing narrow-sense heritability for complex traits. We therefore hypothesized that rare nuclear variants drive a significant proportion of variation in mtDNA-CN. Here we conduct a biobank-scale exome-based study of blood buffy coat.
derived mtDNA-CN in 415,422 individuals of European ancestry.

Material and methods

Sample and variant selection
Data for this study were obtained from the UK Biobank, a resource containing dense genomics, lifestyle, and phenotyping data on ~500,000 volunteers between the ages of 40–69 at recruitment, selectively invited to participate in the prospective study based on the United Kingdom National Health Service (NHS) register and residing within a traveling distance to one of 22 assessment centers in the United Kingdom. Blood collection for blood cell counts and genomic source material extraction was conducted during the recruitment period from 2006–2010. Genotype calls in PLINK format of whole exome sequencing (WES) data for 454,756 samples were accessed on the UK Biobank Research Access Platform (RAP) under application 17731. We started with a total of 26,645,252 variants across autosomal and chromosome X (including targeted and off-target) regions on GRCh38. The samples were processed to remove individuals marked as having withdrawn research consent (45 samples removed), those of self-reported non-White race (25,754 removed using data field 21000, retaining any self-report of White British, White Irish, or any other White background), and those without a computable mtDNA-CN phenotype (13,535 removed) for a total of 415,422 samples. All study participants provided written informed consent and all centers obtained approval from their institutional review boards.

Computation of mtDNA-CN
mtDNA-CN was estimated from DNA derived from the buffy coat, using a combination of whole-exome and genotyping array data, as described previously. In brief, we first extracted reads from 49,997 UK Biobank (UKB) exome sequencing libraries (July 2018 release) and modeled the number of mapped sequencing reads to the reference GRCh38 “chrMT” sequence as a function of total mapped reads, reads to unknown contigs, and decoy sequence reads. Nonlinear effects of total, unknown, and decoy read counts were incorporated using natural splines. The residuals of the adjusted mtDNA mapped read count metric were then regressed onto mitochondrial genotyping probe intensities, available for all 500,000 UKB participants, in separate models for the UK BiLEVE (n = 181 probes) and Affymetrix Axiom (n = 244 probes) genotyping arrays, including in each model nuclear probe intensity-derived principal components to account for technical artifacts. The betas from these models were used to generate fitted values of the mtDNA read count metric in the entire exome sequencing cohort used in this study. Finally, we adjusted the metric for age, sex, and blood cell counts (leukocytes, red blood cells, nucleated red blood cells, platelets, lymphocytes, neutrophils, eosinophils, monocytes, and basophils). All analyses utilized a standardized metric with mean 0 and standard deviation 1. mtDNA-CN was not computed for 13,535 samples that had missing probe intensities or missing cell counts.

Power analysis
We computed the power of the study to detect rare variant associations along the spectrum of expected minor allele frequency (MAF) and effect sizes for hypothesized associations to mtDNA-CN. We based this prediction of the genetic architecture of mtDNA-CN on 129 previously identified genome-wide independent common variants associated with mtDNA-CN, extrapolating a nonlinear power-law regression model fit of MAF and effect sizes of the independent common markers (Figure S1). The R package genpwr v.1.0.4 was used for power calculations.

Single-variant analysis
Single-variant association testing was implemented using the BOLT-LMM mixed-model association testing algorithm, forcing use of a non-infinitesimal mixed-model association test. BOLT-LMM computes a genetic relationship matrix (GRM; or kinship matrix) internally, given a genetic marker set that spans the whole genome. Thus, to account for random effects of genetic relatedness within the BOLT-LMM mixed-model framework, we merged the exome-based genotype calls with 784,256 array-based genotyping calls from the same 415,422 individuals after lifting the coordinates for the array-based variants to hg38 from hg19, and then provided a list of 672,036 (85.7%) of the array-based SNPs, restricted to MAF equal to or greater than 0.001 and SNP missingness equal to or less than 10%, to allow construction of the GRM. The merged variants were filtered for variant- and sample-level missingness greater than 10% (1,193,215 variants removed), monomorphic variants (i.e., those with MAF = 0; 3,286,025 variants removed), and variants with minor allele count less than 6 (19,720,859 variants removed) for a total analysis set of 6,766,699 variants (26.0% of all genotyped exome variants) across autosomes and chromosome X (chrX).

Covariates for analyses included age, sex, 40 principal components (PCs) capturing genetic ancestry, and WES batch (45,200 from the 50,000 initial exomes released and 370,222 from the remaining exomes released). WES batch effects were present because of use of different oligo lots for exome probe capture in the first tranche. Genotypes at 129 independent common autosomal variants and four independent common chrX variants, each previously associated with mtDNA-CN, were included as covariates for autosomal and chrX variant testing, respectively. mtDNA-CN was then regressed onto the genotype using a linear mixed model, including the covariates as fixed effects and a GRM for random effects. For chrX, association testing was stratified by sex, and test statistics and standard errors from the analysis of each sex were combined with METAL. BOLT-LMM internally removed an additional 757 individuals without all covariates available, yielding an analysis set of 414,665 individuals.

Aggregate variant analysis
Aggregate rare variant testing was performed within a mixed-model association testing framework as implemented in the R Bioconductor package GENESIS v.2.21.4. We restricted the analysis to 24,949,430 (93.6%) variants at a MAF equal to or less than 1% across all exome-captured regions. A null mixed model was first fit for mtDNA-CN by regressing the phenotype onto fixed effect covariates of age, sex, genotyping PCs 1–40, and WES batch and including a sparse GRM to account for random effects because of genetic relatedness. The sparse GRM was constructed from kinship estimates provided by the UKB, setting all pairwise relationships with kinship coefficients of less than 0.0442 (third degree or less) to 0. The previously discovered common SNPs were not included as covariates for aggregate testing because of computational cost. Next, for each protein-coding gene, as defined in GENCODE v.38, corresponding to GRCh38.p13, rare-variant burden and adjusted,
asymptotically independent sequence kernel association test (SKAT) p values were computed and then combined in a single set-based mixed model association test (SMMAT) to produce a single p value from a chi-squared distribution with 4 degrees of freedom. Multiple SMMAT tests were carried out for each gene, with overlapping sets of rare variants chosen to test various levels of enrichment of putatively causal variants. Specifically, for each protein-coding gene, an SMMAT test was run for nonsynonymous coding sequence variants passing increasing mutational severity thresholds (all nonsynonymous variants), nonsynonymous variants with a CADD-PHRED score equal to or greater than 18, predicted loss-of-function pLoF variants), and decreasing MAF (≤ 1%, ≤ 0.1%, and ≤ 0.01%), so that, in total, we tested 9 combinations of rare variants, assessing various levels of deleteriousness and allele frequency (Table S1). Each test was required to have a minimum gene-level cumulative allele frequency of 0.01%, but no minimum minor allele count cutoff was set so that singletons and ultra-rare variants with a minor allele count (MAC) of less than 6 were included in the aggregate set. This yielded 153,481 total SMMAT tests. We reasoned that, because the rare variants chosen across different sets for a given gene will have numerous variants in common, the test statistic p values will also be correlated. A correlation matrix of the p values for 13,581 common genes tested across all nine variant sets was used to perform hierarchical clustering, which we conservatively interpreted as outlining four clusters of related p values (Figure S2). Therefore, the number of effective tests was set to the number of unique tested protein coding genes times four, and a correlation-based test p value cutoff was set to 0.05/(18,557 × 4).

Gene-level iterative leave-one-out and conditional variant analyses
For each study-wide significant gene set signal, we used leave-one-out and gene-level conditional analyses to assess the independence of rare variants at that locus. Leave-one-out analysis is a qualitative method that consists of iteratively dropping one variant from a given gene’s variant set and re-computing an SMMAT p value for the locus, producing a measure of the relative contribution of the variant to the gene-based variant set signal. A quantitative linear model of mtDNA-CN was also fit for each study-wide significant gene, testing the conditional independence of each of variant with a MAC equal to or greater than 6 in a single model per gene and adjusting for age, sex, WES batch, and genotyping PCs 1–40.

Gene set enrichment
We downloaded 33,750 gene sets from MitoCarta 3.0, MsigDB v.7.4, and Disgenet v.7.38 For each gene set, we compared means of Z scores from SMMAT tests for genes in and out of the set using a t test assuming equal variance.39 A single aggregate rare variant set, the CADD-PHRED equal to or greater than 18 and MAF equal to or less than 0.001 variant set, was used to obtain Z scores. Because Z scores with extreme values can unduly influence the mean Z score of a gene set, we tested for robustness of resulting sets that were significantly enriched at a Bonferroni-adjusted family-wise error rate (FWER) of 1.5 × 10^-6 (85 gene sets) by excluding extreme-value SMMAT test Z scores (Figure S3), specifically Z scores of TFAM, JAK2, TWNK, and SAMHD1 tests, and re-computing a t test enrichment p value.

Phenome-wide association study (PheWAS)
The R package PheWAS40 was used to collapse summary diagnoses represented as distinct ICD10 codes in UKB field 42170 into binary phenotypes capturing cases and controls for a variety of disease categories. The analysis was restricted to unrelated individuals using the “used.in.pca.calculation” variable26 in resource 531 of the UKB and was restricted to the White British subset of self-reported ancestry using UKB data field 21000. A logistic regression was run for each phenotype as a dependent variable regressed onto the rare variant weighted sum score, adjusting for age, sex, and sequencing center. We used the simpleM41 method to compute the number of effective tests, which accounted for correlated phenotype case and control statuses between clinically related and hierarchical phenotypes. This yielded 1,530 effective tests, which was used as the denominator in the Bonferroni multiple testing correction. The number of effective tests times the number of genes (i.e., p ≤ 3.0 × 10^-6) was used as the Bonferroni multiple test correction denominator for the pheWAS of rare variant carrier status in individual genes. The R package MendelianRandomization42 was used to obtain the inverse variance weighted causal point estimate between mtDNA-CN rare variants in SAMHD1 and the phenotype for breast cancer risk.

All statistical analyses were performed in R 4.0.2.

Results
Sample characteristics
The current study included 415,422 self-reported White individuals (54.2% female) with a mean age of 56.8 (SD = 8.03) years enrolled in the UKB. mtDNA-CN was estimated from DNA derived from blood buffy coat12 and was adjusted for blood cell counts including immune cells (leukocytes, platelets, lymphocytes, neutrophils, eosinophils, monocytes, and basophils) to account for the influence of immune cells on mtDNA-CN.27

Single variants significantly associated with mtDNA-CN
We performed linear mixed-model association testing29 of single-nucleotide variants (SNVs) and small insertions or deletions (indels) with a MAC equal to or greater than 6 in 414,665 UKB exomes of self-reported Whites, adjusting for fixed effects of age, sex, sequencing batch, and random effects because of genetic relatedness and ancestry. We also included genotypes at 129 previously discovered autosomal common variant GWAS loci2 as fixed effects to condition the results on independent genome-wide signals from common single variants associated with mtDNA-CN. We identified 11 SNVs associated with mtDNA-CN at a Bonferroni-corrected genome-wide significance (p ≤ 7.4 × 10^-9) (Table 1). The most significant SNVs were a pair of rare variants separated by 1.2 Mb in SAMHD1 (g.36893060C>T [GenBank: NC_000020.11], p.Asp585Asn, beta = +0.61 SD [SE = 0.06], p = 4.2 × 10^-28, MAC = 304) and SPAG4 (g.35619007C>G [GenBank: NC_000020.11], beta = +0.55 SD [SE = 0.06], p = 6.4 × 10^-22, MAC = 288) across chromosome 20q11.22–q11.23 (Figure 1A).

Conditioning on the common variant signal allowed us to uncover an independent rare non-coding variant within intron one of SIRPA (g.1895355G>C [GenBank: NC_
0.0029, p = 0.35; Figure S7).

Finally, we tested the robustness of the genome-wide significant single-variant associations to removal of cell count outlier individuals (n = 4,077), finding no overall significant difference in effect estimates (paired t test p = 0.68; Figure S8). However, although significant with and without outlier removal, the missense variant in JAK2 had a significantly stronger effect estimate in cell count outlier individuals (beta = 0.40 [SE = 0.08] versus beta = 0.83 [SE = 0.12], p for difference = 1.5 x 10^-4), and a large proportion of variant carriers are cell count outliers (117/274 individuals, 42.7%) because of prevalent myeloproliferative disease in these individuals.

**Rare, ancestral haplotype on 20q11 significantly associated with mtDNA-CN**

Given that the rare single variant associations in SAMHD1 and SPAG4 are in close physical proximity and have a similar MAC, we investigated the independence of the variant effects on mtDNA-CN. The rare SNVs in SPAG4 (MAC 288, p = 6.4 x 10^-22) and SAMHD1 (MAC 304, p = 4.2 x 10^-28) were present in a shared set of 262 individuals (Figure 1A). The association with mtDNA-CN, however, was only significant among carriers of both variants (p = 9.94 x 10^-22) and those with only the SAMHD1 variant (40 heterozygotes, p = 0.002) (Figure S9). Conditional analysis including both SNPs in the same model.
confirmed that the association with mtDNA-CN was driven by the \textit{SAMHD1} variant (conditional $p = 1.1 \times 10^{-6}$, \textit{SPAG4} conditional $p = 0.59$).

These data are consistent with the presence of a rare ancestral haplotype segregating in this population between \textit{SPAG4} and \textit{SAMHD1} that is associated with mtDNA-CN, with the association with mtDNA-CN at \textit{SPAG4} being solely due to its linkage disequilibrium with \textit{SAMHD1}. The 1.2-Mb region between \textit{SPAG4} and \textit{SAMHD1} lies in a previously described recombination desert (Figure S10), and the surrounding DNA on 20q11 is therefore identical by descent in the subset of UKB participants who share this haplotype.

\textbf{Gene-based groups of rare variants significantly associated with mtDNA-CN}

We next implemented a linear mixed-model aggregate rare variant association testing framework to test for association between sets of rare variants in protein-coding genes and mtDNA-CN. A null mixed-effects model was fit, accounting for random effects of genetic kinship and fixed effects of age, sex, sequencing batch, and genotype-derived PCs of ancestry. Because of a lack of priors on true causality that can be used to identify the ideal subset of rare variants to be included in a gene-based test, for each gene we tested 9 sets of nonsynonymous protein-coding sequence variant combinations at increasing mutational severity thresholds (all nonsynonymous variants, nonsynonymous variants with CADD-PHRED score $R \geq 18$, pLoF variants), and decreasing MAF ($\% \geq 1\%$, $\% \leq 0.1\%$, and $\% \leq 0.01\%$). Eight genes were associated with mtDNA-CN at a study-wide significance cutoff of $p \leq 6.7 \times 10^{-7}$ (Table 2). Three additional genes (\textit{POLRMT}, $p = 1.1 \times 10^{-6}$; \textit{CAVIN2}, $p = 1.9 \times 10^{-6}$; \textit{MGME1}, $p = 2.0 \times 10^{-6}$) had marginal significance (Table S3), of which two of three (\textit{POLRMT} and \textit{MGME1}) have previously been linked to mitochondrial function. \textit{POLRMT} encodes the RNA polymerase for mtDNA and is part of the mitochondrial transcription initiation complex, which includes TFAM and TFB2M, that is required for basal transcription of mtDNA. \textit{MGME1} is a mtDNA maintenance exonuclease, and loss-of-function mutations within \textit{MGME1} have been linked...
with autosomal recessive mtDNA depletion syndrome 11 (MTDPS11 [MIM: 615084]).

For each gene, we used an iterative leave-one-out analysis to test the hypothesis that a subset of variants at each locus is gene-level conditionally independent and drives most of the association signal. Across eight genes, we identified 18 gene-level Bonferroni-significant conditionally independent associations (Table S4), including 4 missense mutations in the most significantly associated gene, SAMHD1 (gene-based test p = 4.3 × 10^{-45}; Figures 1B and 1C). Conditionally independent variants had consistent effect directions within a gene, in line with the recent observation of long allelic series of rare coding variants in a large-scale exome study in the UKB and suggesting that independent alterations to distinct protein residues within the same gene affected mtDNA-CN in a consistent manner. Three genes (CHEK2, POLRMT, and CAVIN2) had no individual variant that was significantly associated after gene-wide multiple-testing correction, suggesting that a multitude of small effects across the genes drove the gene set signal at these loci.

Most loci with conditionally independent effects had trait-increasing alleles, with only TWNK (SMMAT test p = 1.1 × 10^{-30}) and TFAM (SMMAT test p = 4.3 × 10^{-15}) harboring evidence of multiple independent mtDNA-CN-lowering effect alleles, consistent with their role and localization in the mtDNA nucleoid and their implication in autosomal recessive Mendelian mtDNA depletion syndromes. Only TWNK and TFAM harbored ultra-rare (MAC < 6) nonsynonymous variants that had significant association with mtDNA-CN. These MAC less than 6 variants, coded as a single synthetic allele, were associated with lower mtDNA-CN in TWNK (beta = −0.25 SD [SE = 0.04]), p = 2.17 × 10^{-10}) and TFAM (beta = −0.37 SD [SE = 0.06], p = 8.3 × 10^{-9}).

**Table 2.** Aggregate gene-based rare variants associated with mtDNA-CN at study-wide significance (p ≤ 6.7 × 10^{-7}).

| Number of Qualifying Variant Sites | Number of Non-Reference Alleles | Number of Individuals with Non-Reference Alleles | Cumulative Allele Frequency | P-value | P-value | P-value | Q-value | Gene | Variant Set |
|-----------------------------------|--------------------------------|-----------------------------------------------|----------------------------|---------|---------|---------|---------|------|-------------|
| 258                               | 1,747                          | 1,739                                         | 0.0021                     | 4.25 × 10^{-37} | 9.32 × 10^{-11} | 4.27 × 10^{-45} | 6.75 × 10^{-40} | SAMHD1 | Nonsynonymous, CADD-PHRED ≥ 18, MAF ≤ 0.001 |
| 278                               | 1,304                          | 1,296                                         | 0.0016                     | 2.20 × 10^{-18} | 6.70 × 10^{-15} | 1.09 × 10^{-30} | 3.46 × 10^{-26} | TWNK   | Nonsynonymous, CADD-PHRED ≥ 18, MAF ≤ 0.0001 |
| 167                               | 1,873                          | 1,870                                         | 0.0023                     | 2.74 × 10^{-10} | 4.12 × 10^{-7}  | 4.27 × 10^{-15} | 5.62 × 10^{-11} | TFAM   | Nonsynonymous, MAF ≤ 0.01 |
| 366                               | 4,853                          | 4,796                                         | 0.0058                     | 2.84 × 10^{-4}  | 1.99 × 10^{-12} | 2.04 × 10^{-14} | 2.48 × 10^{-10} | JAK2   | Nonsynonymous, CADD-PHRED ≥ 18, MAF ≤ 0.0001 |
| 969                               | 56,902                         | 41,702                                        | 0.0685                     | 8.75 × 10^{-9}  | 1.56 × 10^{-4}  | 3.86 × 10^{-11} | 3.21 × 10^{-7}  | LY75   | Nonsynonymous, MAF ≤ 0.01 |
| 1,137                             | 58,911                         | 43,513                                        | 0.0709                     | 2.96 × 10^{-8}  | 2.21 × 10^{-4}  | 1.75 × 10^{-10} | 1.32 × 10^{-6}  | LY75, CD302 | Nonsynonymous, MAF ≤ 0.01 |
| 309                               | 13,566                         | 13,419                                        | 0.0163                     | 2.90 × 10^{-5}  | 1.27 × 10^{-5}  | 8.41 × 10^{-9}  | 5.11 × 10^{-5}  | CLPX   | Nonsynonymous, MAF ≤ 0.01 |
| 131                               | 18,663                         | 18,388                                        | 0.0241                     | 4.19 × 10^{-6}  | 5.33 × 10^{-4}  | 4.68 × 10^{-8}  | 2.55 × 10^{-4}  | AK2    | Nonsynonymous, CADD-PHRED ≥ 18, MAF ≤ 0.01 |
| 457                               | 11,048                         | 10,947                                        | 0.0133                     | 2.58 × 10^{-4}  | 8.75 × 10^{-3}  | 4.19 × 10^{-7}  | 2.14 × 10^{-3}  | CHEK2  | Nonsynonymous, MAF ≤ 0.01 |

**a**Cumulative allele frequency defined as summation of non-reference allele frequencies for all qualifying variant sites in gene.

**b**SMMAT p-value combines burden test and SKAT in a single p-value from chi-squared distribution with 4 degrees of freedom.

**c**LY75 / LY75-CD302 represent a single naturally occurring readthrough locus producing a single fusion protein.

Rare variants for mtDNA-CN significantly associated with Mendelian mtDNA depletion syndromes

To test the hypothesis that mtDNA-CN-associated rare variants are enriched in known causal mtDNA depletion syndrome genes, we tested whether the 20 known causal depletion syndrome genes have excess association signal in our data. Eight of 20 genes were nominally associated at p ≤ 0.05 (3 genes at Bonferroni-adjusted p ≤ 0.05, accounting for 20 tests), consistent with highly significant enrichment (exact binomial test p = 2.9 × 10^{-6}; Table S5).

**Gene set enrichment analysis**

We next tested for enrichment of aggregate gene-based rare variant signal within each of 33,750 annotated gene sets compiled from MsigDb,37 Mito Carta,4 and Disgenenet38 under the hypothesis that loci harboring rare variants for mtDNA-CN would implicate functionally relevant gene...
Finally, a gene set composed of common variant-implied genes in the set [cated loci12 was also significantly enriched for rare variant

Figure 2. mtDNA-CN-associated rare variants in SAMHD1 exhibit a linear dose-response relationship with breast cancer risk

Effect estimates for gene-level conditionally independent associations with mtDNA-CN from four uncorrelated rare variants in SAMHD1 are plotted against the corresponding association betas for disease phecode 174, representing breast cancer risk (n = 223,098 females, k = 17,517 cases; Table S7). The slope of the blue line illustrates the inverse-variance weighted causal estimate, and bars around points represent 95% confidence intervals.

Disease PheWAS
To test the hypothesis that rare genetic variation associated with mtDNA-CN highlights causal pathways between mtDNA-CN and disease risk, we collapsed summary diagnoses from inpatient hospital records represented by distinct ICD10 codes for 314,201 unrelated individuals of White British ancestry in the UKB into 1,623 broad phecodes. We then tested for association in a pheWAS between the phecodes and rare variant carrier status, represented as a mtDNA-CN-associated rare variant summed score weighted by effect sizes, reasoning that rare variants affecting mtDNA-CN act additively to modulate risk of a common downstream phenotype. The weighted sum score was restricted to variants of an effect size larger than ±0.3 SD on mtDNA-CN and included a synthetic allele of ultra-rare variants with MAC less than 6 in TFAM that had an effect size of less than −0.3 SD. PheWAS of the mtDNA-CN rare variant score yielded Bonferroni-significant associations with myeloproliferative disease (phecode 200) and myeloid leukemia (phecode 204.2). To assess whether these phecode associations were driven by rare variants at a single gene, we removed the variant at the JAK2 locus, which was individually strongly associated with the same phecodes (MAC = 284, odds ratio [OR] = 822.4 [95% confidence interval [CI] = 821.0–823.9], t statistic = 36.9, p = 3.2 × 10−299 for phecode 200; OR = 166.4 [95% CI = 164.6–168.2], t statistic = 17.1, p = 8.5 × 10−66 for phecode 204.2) and noted that there were no significant associations with any disease phecodes after removing the JAK2 variant from the weighted sum score, indicating that the association of the rare variant score to myeloproliferative disease was driven by the rare variant at the JAK2 locus. We further ran pheWAS separately for each gene implicated by rare variation and found that carrier status of conditionally independent mtDNA-CN-associated SAMHD1 rare variants (MAC ≥ 6) was marginally associated with increased risk of breast cancer (phecode 174, k = 15,958 cases, OR = 2.1 [95% CI = 1.5–2.9], p = 1.9 × 10−5). To assess whether any single conditionally independent variant at the SAMHD1 locus drove this association, we tested each variant separately for breast cancer risk. There was a linear dose-response relationship with breast cancer risk for the four conditionally independent, uncorrelated mtDNA-CN associated variants at the SAMHD1 locus (Figure 2; inverse-variance weighted Mendelian randomization causal point estimate = 1.29 [SE = 0.253], p = 3.1 × 10−7; Table S7).

Discussion
We conducted an exome-based study of genetic variation associated with mtDNA-CN in 415,422 individuals from the UKB. We report 24 independent signals from 17 loci, including nine SNVs and one rare noncoding variant identified as study-wide independent. Using an iterative leav-
one-out approach, we fine-mapped 14 gene-level conditionally independent effects. Rare variants were enriched in loci implicated in Mendelian mtDNA depletion syndromes, and these variants further implicated core processes involved with mtDNA replication, nucleoid structure formation, and maintenance. While the manuscript was in preparation, Chong et al. identified SAMHD1 from 147,740 samples as the only exome-based association signal for mtDNA-CN estimated using a different approach, which provides additional evidence for our observed association of SAMHD1 associated with mtDNA-CN. In part because of the approximately three times larger sample size, this study found multiple additional loci. Combined with gene-based tests, we also identify 10 loci not detectable by common variant GWASs (MAF < 0.002), showcasing the added value of rare variants for discovery of loci contributing to the genetic architecture of mtDNA-CN.

In contrast to the expectation that physically distant rare variants are likely to be in linkage equilibrium, we identified a pair of rare variants residing in a unique shared haplotype associated with mtDNA-CN. The variants in SPAG4 and SAMHD1 were separated by 1.2 Mb on 20q11 and formed a single rare, shared haplotype in 262 individuals (haplotype frequency 3.2 × 10⁻⁴). The presence of a shared ancestral haplotype is supported by a similar MAC pattern for the rare variants in SPAG4 and SAMHD1 within the publicly available TOPMed Freeze 8 data (MAC = 21 and MAC = 25 for g.35619007C>G [GenBank: NC_000020.11] and g.36893060C>T [GenBank: NC_000020.11], respectively), suggesting that this haplotype is older than 5–10 generations because the TOPMed cohort consists of more than 80% United States residents who experienced a bottleneck during the migration to America. The shared haplotype lies in a known region of very low recombination on 20q, suggesting that the persistence of the identical-by-descent segment in this population bearing this haplotype can be attributed to the low recombination rate. This finding warrants further inspection of rare haplotype sharing affecting complex traits in large biobank-scale sequencing datasets.

The association with mtDNA-CN by the rare haplotype was solely driven by the SAMHD1 mutation (p.Asps85Asn [GenBank: NP_056289.2]), which offers insight into the molecular mechanism behind how variation in SAMHD1 leads to increased mtDNA-CN. SAMHD1 encodes a deoxynucleoside triphosphate triphosphohydrolase, a host restriction factor enzyme that depletes dNTP pools and aids in the viral defense response. The deoxynucleoside products of the enzymatic hydrolysis by SAMHD1 are transported into the mitochondria and salvaged to dNTPs. Although an overabundance of dNTPs could indicate a putative pathway by which het erozygous mutations within SAMHD1 lead to increased mtDNA-CN, because one determinant of mtDNA-CN is nucleotide metabolism, we found that the p.Asps85Asn variant falls within a region of SAMHD1 deleted in autosomal recessive Aicardi-Goutières syndrome 5 (AGS5 [MIM: 612952]). In particular, a homozygous p.Gln548Ter variant in the same region of the gene found in an individual with SAMHD1 deficiency has been demonstrated to influence SAMHD1’s alternative role in degrading nascent single-stranded DNA (ssDNA) at stalled replication forks but not influence dNTPase activity, suggesting that the conditionally independent variant in our study may have an alternative molecular mechanism revolving around failure of ssDNA degradation.

SAMHD1’s role as a regulator of the innate immune response and a key part of the replication stress response also supports its role in cancer and allows consideration of the hypothesis that mtDNA-CN may be causally associated with cancer. Conditional analysis of rare variants within SAMHD1 revealed independent mtDNA-CN-increasing effects at four sites within SAMHD1, and there was a linear dose-response relationship between the four variants’ effects on mtDNA-CN and on breast cancer risk. These data support a mechanistic link between genetic variation at the SAMHD1 locus on mtDNA-CN and on breast cancer. Importantly, however, we did not find a disease pheWAS signal for breast cancer among individuals carrying non-SAMHD1 mtDNA-CN-associated rare variants, suggesting mtDNA-CN is not causal for breast cancer risk but that the mechanism that impacts mtDNA-CN because of SAMHD1 deficiency is having a proportional impact on breast cancer risk at the SAMHD1 locus.

We also found a subset of other genes harboring mtDNA-CN-associated rare variants in our study (JAK2, TOP3A, CHEK2, and SIRPA) that are pleiotropic for cancer phenotypes. The missense mutation at the JAK2 locus, p.Val1617Ile, is a dominant gain-of-function mutation that is causal for myeloproliferative disease, possibly a result of clonal hematopoiesis of intermediate potential (CHIP). Biallelic mutations at TOP3A have been associated with a Bloom syndrome-like disorder, a rare disorder characterized by predisposition to cancer, albeit with additional mitochondrial dysfunction consistent with the TOP3A function in mtDNA decatenation. CHEK2 is a tumor suppressor gene that encodes a checkpoint protein kinase activated because of DNA damage and controls cell-cycle arrest, and mutations in CHEK2 have been associated with autosomal dominant forms of breast, colorectal, and prostate cancer as part of familial Li-Fraumeni syndrome (LFS [MIM: 151623]). SIRPA, a gene previously identified in the common variant GWAS for mtDNA-CN, encodes a cell surface receptor for CD47 involved in prevention of auto-phagocytosis of myeloid cells that is frequently downregulated in a variety of cancers. Although we found individual pleiotropic associations with cancer at loci implicated by rare variant associations with mtDNA-CN, combined, the mtDNA-CN rare variant weighted sum score was not associated with any cancer phenotypes, lending credence to a competing hypothesis that mtDNA-CN is significantly dysregulated in a variety of
cancers and that it is downstream of pathways leading to cancer.

Finally, studying rare variants associated with mtDNA-CN allowed us to pinpoint functional variants in known genes implicated in autosomal recessive Mendelian mtDNA depletion syndromes (TNWK and TFAM). The most significant TNWK variant decreasing mtDNA-CN was a missense mutation at g.100993257G>A (GenBank: NC_000010.11) (MAC = 48, beta = −1.09 SD [SE = 0.14], p = 1.4 × 10−14), a variant previously reported in two affected siblings with a diagnosis of autosomal recessive Perrault syndrome 565 (PRLTS5 [MIM: 616138]) who each harbored an additional compound heterozygous mutation in TNWK and whose clinical presentation included muscle weakness and atrophy reminiscent of the broader MTDPS7 (MIM: 271245), the autosomal recessive disorder caused by mutations in TNWK. TFAM encodes a transcription factor part of the initiation complex required for basal transcription of mtDNA and is a key component of mtDNA nucleoids, and mutations within TFAM have previously been linked with mtDNA depletion syndrome 15 (MTDPS15 [MIM: 617156]). Two variants were gene-level conditionally independent at TFAM at g.5838876C>G (GenBank: NC_000010.11) (MAC = 320, amino acid [aa] residue Q100X, beta = −0.27 SD [SE = 0.058], p = 3.0 × 10−6) and at g.58394946A>G (GenBank: NC_000010.11) (MAC = 81, aa residue K173E, beta = −0.49 SD [SE = 0.11], p = 1.3 × 10−5), which each fall within the experimentally validated DNA-binding domains for TFAM: the high-mobility group (HMG) box 1 (residues 50–118) and HMG box 2 (residues 155–219), respectively. The conditional independence of the decreasing mtDNA-CN effects from variants in each of the DNA binding domains within TFAM confirms that both residues are critical for proper functioning of the protein.

Several limitations exist for this study. First, the SNP-based heritability of the measure of mtDNA-CN used in this study is 7.0%−7.4%, as estimated by common variants, which is in contrast to heritability estimates of 54%−65% from previous family-based studies for mtDNA-CN. Notwithstanding artificial inflation of heritability estimates from twin- and family-based studies, this gap suggests that, among other sources of genetic variation that could contribute to the missing heritability, including rare variants, an imperfect phenotypic measurement affects the association test power of common and rare variant GWASs. mtDNA-CN is “noisy” because of confounding from platelet counts and other types of cells, and although we account for these cell counts in our measurement, hidden sources of variability were not considered that might impact association test power. Analogously, blood mtDNA-CN has other determinants beyond mitochondrial function, which limits interpretation of the rare genetic variants impacting the phenotype.

Second, despite the large sample size of the study, the power of the study to detect rare variant associations for mtDNA-CN is still limited. We investigated this deficit by extrapolating a power-law fit from the effect sizes and MAF of the 129 autosomal common variants for mtDNA-CN and estimated power along the spectrum of MAF and effect sizes expected for a presumed rare variant genetic architecture. For a study cohort size of 414,665 individuals and controlling type I error at a Bonferroni-corrected alpha of 7.4 × 10−9, the present study is underpowered to detect rare single-variant associations with mtDNA-CN along the predicted effect sizes for rare variants with MAF less than 1% (Figure S1). Finally, the study is limited to self-reported White individuals because of a lack of sample size for other genetic ancestry groups.

We conducted the largest-to-date study of rare variation impacting mtDNA-CN, surveying over 400,000 individuals sourced from a single tissue and collection method. Using an association testing framework that tested single and gene-based groups of variants, 17 genes were identified. These genes highlight the biological underpinnings of mtDNA-CN and mitochondrial function, including enrichment in genes associated with Mendelian mtDNA depletion syndrome genes. Combined, these data highlight the importance of studying rare genetic variation in the study of the complex mtDNA-CN phenotype.

Data and code availability

WES data for the UKB is available for access through the UKB RAP with an approved UKB application. Summary statistics for mtDNA-CN rare-variant association tests are available for download at https://www.arkinglab.org/resources/. Scripts and code for generation of association statistics and additional statistical analyses are available at https://github.com/ArkingLab/mtDNA-CN-ExWAS.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2022.100147.

Acknowledgments

This study was supported by the National Institutes of Health under grants R01HL13573 and R01HL144569. V.P. was funded by the Burroughs Wellcome Fund. The research utilized the UK Biobank Resource under application number 17731. The authors thank Charles Newcomb for valuable input and technical assistance.

Author contributions

D.E.A. and V.P. conceptualized the study, designed the experiments, and edited the manuscript. V.P. conducted the experiments, performed formal statistical analyses, including design, creation and execution of scripts and code, and wrote and edited the manuscript. W.S., C.S., S.Y., and J.L. conducted the experiments, curated data, and provided technical and statistical analysis support. D.E.A., E.G., and N.P. designed experiments, provided research supervision, and edited the manuscript. All authors reviewed the manuscript.
Mitochondrial DNA content: its genetic heritability and association with renal cell carcinoma. J. Natl. Cancer Inst. 100, 1104–1112.

25. Wainschtein, P., Jain, D., Zheng, Z., TOPMed Anthropometry Working Group, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Adrienne Cupples, L., Shadayab, A.H., McKnight, B., Shoemaker, B.M., Mitchell, B.D., Psaty, B.M., et al. (2022). Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. Nat. Genet. 54, 263–273.

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

27. Picard, M. (2021). Blood mitochondrial DNA copy number: what are we counting? Mitochondrion 60, 1–11.

28. Moore, C.M., Jacobson, S.A., and Fingerlin, T.E. (2019). Power and sample size calculations for genetic association studies in the presence of genetic model misspecification. Hum. Hered. 84, 256–271.

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

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

31. Gogarten, S.M., Sofer, T., Chen, H., Yu, C., Brody, J.A., Thornton, T.A., Rice, K.M., and Conomos, M.P. (2019). Genetic association testing using the GENESIS R/Bioconductor package. Bioinformatics 35, 5346–5348.

32. Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J.E., Mudge, J.M., Sisu, C., Wright, J.C., Armstrong, J., Barnes, I., Berry, A., et al. (2021). Gencode 2021. Nucleic Acids Res. 49, D916–D923.

33. Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. Am. J. Hum. Genet. 95, 5–23.

34. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. 89, 82–93.

35. Chen, H., et al. (2019). Efficient variant set mixed model association tests for Continuous and binary traits in large-scale whole-genome sequencing studies. Am. J. Hum. Genet. 104, 260–274.

36. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 47, D886–D894.

37. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics 27, 1739–1740.

38. Piñero, J., Ramirez-Anguita, J.M., Saúch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., and Furlong, L.I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 48, D845–D855.

39. Irizarry, R.A., Wang, C., Zhou, Y., and Speed, T.P. (2009). Gene set enrichment analysis made Simple. Stat. Methods Med. Res. 18, 565–575.

40. Carroll, R.J., Bastarache, L., and Denny, J.C. (2014). R PhEWas: data analysis and plotting tools for phenotype-wide association studies in the R environment. Bioinformatics 30, 2375–2376.

41. Gao, X., Starmer, J., and Martin, E.R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genet. Epidemiol. 32, 361–369.

42. Yavorska, O.O., and Burgess, S. (2017). MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. Int. J. Epidemiol. 46, 1734–1739.

43. Kralovics, R., Passamonti, F., Buser, A.S., Teo, S.S., Tiedt, R., Passweg, J.R., Tchelli, A., Cazzola, M., and Skoda, R.C. (2005). A gain-of-function mutation of JAK2 in myeloproliferative disorders. N. Engl. J. Med. 352, 1779–1790.

44. Deloukas, P., Matthews, L.H., Ashurst, J., Burton, J., Gilbert, J.G., Jones, M., Stavrides, G., Almeida, J.P., Babbage, A.K., Baggley, C.L., et al. (2001). The DNA sequence and comparative analysis of human chromosome 20. Nature 414, 865–871.

45. Hillen, H.S., Morozov, Y.I., Sarfallah, A., Temiakov, D., and Cramer, P. (2017). Structural basis of mitochondrial transcription initiation. Cell 171, 1072–1080.

46. Kornblum, C., Nicholls, T.J., Haack, T.B., Schön, S., Pvee, V., Danhauer, K., Hallmann, K., Zsurka, G., Rorbach, J., Iuso, A., Wieland, T., et al. (2013). Loss-of-function mutations in MEG1 impair mtDNA replication and cause multisystemic mitochondrial disease. Nat. Genet. 45, 214–219.

47. Barton, A.R., Sherman, M.A., Mukamel, R.E., and Loh, P.-R. (2021). Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. Nat. Genet. 53, 1260–1269.

48. OMIM Phenotypic Series - PS603041. 2012, https://omim.org/phenotypicSeries/PS603041.

49. Chong, M., Mohammadi-Shemirani, P., Perrot, N., Nelson, W., Morton, R., Narula, S., Lali, R., Khan, I., Khan, M., Judge, C., et al. (2022). GWAS and ExWAS of blood mitochondrial DNA copy number identifies 71 loci and highlights a potential causal role in dementia. Elife 11, e70382.

50. Li, B., Liu, D.J., and Leal, S.M. (2013). Identifying rare variants associated with complex traits via sequencing. Curr. Protoc. Hum. Genet. Chapter 1, Unit–26.

51. Turkmen, A., and Lin, S. (2017). Are rare variants really independent? Genet. Epidemiol. 41, 363–371.

52. NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature 590, 290–299.

53. Franzolin, E., Pontarini, G., Rampazzo, C., Mazzarini, C., Ferraro, P., Palumbo, E., Reichard, P., and Bianchi, V. (2013). The deoxynucleotide triphosphohydrolase SAMHD1 is a major regulator of DNA precursor pools in mammalian cells. Proc Natl Acad Sci USA 110, 14272–14277.

54. Franzolin, E., Salata, C., Bianchi, V., and Rampazzo, C. (2015). The deoxynucleotide triphosphate triphosphohydrolase activity of SAMHD1 protein contributes to the mitochondrial DNA depletion associated with genetic deficiency of deoxynucleosine kinase. J. Biol. Chem. 290, 25986–25996.

55. Coquel, F., Silva, M.J., Técher, H., Zadorozhny, K., Sharma, S., Nieminszycz, J., Mettling, C., Dardillat, E., Barthe, A., et al. (2018). SAMHD1 acts at stalled replication forks to prevent interferon induction. Nature 557, 57–61.
56. Clifford, R., Louis, T., Robbe, P., Ackroyd, S., Burns, A., Timbs, A.T., Wright Colopy, G., Dreu, H., Sigaux, F., Judde, J.G., et al. (2014). SAMHD1 is mutated recurrently in chronic lymphocytic leukemia and is involved in response to DNA damage. Blood 123, 1021–1031.

57. Rentoft, M., Lindell, K., Tran, P., Chabes, A.L., Buckland, R.J., Watt, D.L., Marjavaara, L., Nilsson, A.K., Melin, B., Trygg, J., Johansson, E., and Chabes, A. (2016). Heterozygous colon cancer-associated mutations of SAMHD1 have functional significance. Proc Natl Acad Sci USA 113, 4723–4728.

58. Wang, J., Lu, F., Shen, X.-Y., Wu, Y., and Zhao, L. (2014). SAMHD1 is down regulated in lung cancer by methylation and inhibits tumor cell proliferation. Biochem. Biophys. Res. Commun. 455, 229–233.

59. Genovese, G., Kähler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., Bakhoum, S.F., Chambert, K., Mick, E., Neale, B.M., et al. (2014). Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N. Engl. J. Med. 371, 2477–2487.

60. Jaiswal, S., Natarajan, P., Silver, A.J., Gibson, C.J., Bick, A.G., Shvartz, E., McConkey, M., Gupta, N., Gabriel, S., Ardissino, D., et al. (2017). Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. N. Engl. J. Med. 377, 111–121.

61. Martin, C.-A., Sarlós, K., Logan, C.V., Thakur, R.S., Parry, D.A., Bizard, A.H., Leitch, A., Cleal, L., Ali, N.S., Al-Owain, M.A., et al. (2018). Mutations in TOP3A cause a Bloom syndrome-like disorder. Am. J. Hum. Genet. 103, 221–231.

62. Nicholls, T.J., Nadalutti, C.A., Motori, E., Sommerville, E.W., Gorman, G.S., Basu, S., Hoberg, E., Turnbull, D.M., Chinnery, P.F., Larsson, N.G., et al. (2018). Topoisomerase 3α Is Required for Decatenation and Segregation of Human mtDNA. Mol. Cell 69, 9–66.

63. Bell, D.W., Varley, J.M., Szydlo, T.E., Kang, D.H., Wahrer, D.C., Shannon, K.E., Lubratovich, M., Verselis, S.J., Isselbacher, K.J., Fraumeni, J.F., et al. (1999). Heterozygous germ line hCHK2 mutations in Li-Fraumeni syndrome. Science 286, 2528–2531.

64. Takahashi, S. (2018). Molecular functions of SIRPz and its role in cancer. Biomed. Rep. 9, 3–7.

65. Oldak, M., Ozieblo, D., Pollak, A., Stepniak, I., Lazniewski, M., Lechowicz, U., Kochanek, K., Furmanek, M., Tacikowska, G., Plewczynski, D., et al. (2017). Novel neuro-audiological findings and further evidence for TWNK involvement in Perrault syndrome. J. Transl. Med. 15, 25.

66. Vozárková, V., et al. (2020). Mitochondrial HMG-box containing proteins: from Biochemical Properties to the roles in human diseases. Biomolecules 10, 1193.

67. Campbell, C.T., Kolesar, J.E., and Kaufman, B.A. (2012). Mitochondrial transcription factor A regulates mitochondrial transcription initiation, DNA packaging, and genome copy number. Biochim Biophys Acta 1819, 921–929.

68. Larsson, N.-G., Wang, J., Wilhelmsson, H., Oldfors, A., Rustin, P., Lewandowski, M., Barsh, G.S., and Clayton, D.A. (1998). Mitochondrial transcription factor A is necessary for mtDNA maintenance and embryogenesis in mice. Nat. Genet. 18, 231–236.

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

70. Mayhew, A.J., and Meyre, D. (2017). Assessing the heritability of complex traits in humans: Methodological challenges and opportunities. Curr. Genomics 18, 332–340.

71. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature 461, 747–753.