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,1,2 John Lane,4 Eliseo Guallar,5 Nathan Pankratz,4 Dan E. Arking1,*

1McKusick-Nathans Institute, Department of Genetic Medicine, Johns Hopkins University School of Medicine
2Predoctoral Program in Human Genetics, Johns Hopkins University School of Medicine
3Maryland Genetics Epidemiology and Medicine Training Program, Johns Hopkins Bloomberg School of Public Health
4Department of Laboratory Medicine and Pathology, University of Minnesota Medical School
5Departments of Epidemiology and Medicine, and Welch Center for Prevention, Epidemiology and Clinical Research, Johns Hopkins Bloomberg School of Public Health

* Correspondence: arking@jhmi.edu

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

Inter-individual variation in the number of copies of the mitochondrial genome, termed 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, both 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 mitochondrial DNA depletion syndrome genes (mtDNA helicase TWNK, $p=5.7 \times 10^{-29}$; mitochondrial transcription factor TFAM, $p=4.3 \times 10^{-13}$; mtDNA maintenance exonuclease MGME1, $p=1.3 \times 10^{-6}$) and the V617F dominant gain-of-function mutation in the tyrosine kinase JAK2 ($p=7.1 \times 10^{-17}$) associated with myeloproliferative disease. Novel genes included the ATP-dependent protease CLPX ($p=9.9 \times 10^{-9}$) involved with mitochondrial proteome quality and the mitochondrial adenylate kinase AK2 ($p=5.3 \times 10^{-8}$) involved with hematopoiesis. 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 with 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 further implicated core processes in mtDNA replication, nucleoid structure formation, and maintenance. Together, these data indicate 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\(^1\), and in the regulation of innate immunity\(^2,3\).

Most of mitochondrial form and function, including factors required to construct the organelle\(^4\), is encoded within the nuclear genome, yet mitochondria have evolved\(^5,6\) 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, both at levels considered heteroplasmic and homoplasmic, contribute to disease risk\(^7\) including Mendelian mtDNA depletion syndromes\(^8\) and aging-related disease\(^9\).

Unlike the fixed content of nuclear DNA, mtDNA ploidy is variable, both within cells and across tissues, individuals, and cohorts\(^10,11\). Indeed, a portion of inter-individual variation in the number of mitochondrial genome copies (mtDNA-CN) reflects mitochondrial function and is heritable under nuclear genetic control via common polygenic variation\(^12,13\). As a minimally invasive biomarker, mtDNA-CN is associated with several aging-relevant traits including cardiovascular disease\(^11,14–16\), frailty and all-cause mortality\(^17\), neurodegeneration\(^18–20\), and cancer\(^21,22\).

Genome-wide association studies (GWAS) for mtDNA-CN have revealed mtDNA-CN to be a complex phenotype, with SNP-based heritability estimates between 7-8%. Family-based studies, however, have estimated broad sense heritability for mtDNA-CN at 54-65\(^{23,24}\), and variance explained by common variants in a non-discovery cohort was 1-2\(^{12}\). Rare genetic variants, especially mutations within protein-coding regions of the nuclear genome, collectively account for a substantial proportion of missing narrow-sense heritability for complex traits\(^25\). We therefore hypothesized that rare nuclear variants drive a significant proportion of variation in mtDNA-CN. Herein, we conduct a biobank-scale exome-based study of blood buffy-coat-derived mtDNA-CN in 415,422 individuals of European ancestry.
Materials and Methods

Sample and variant selection

Data for this study were obtained from the UK Biobank\textsuperscript{26}, a resource containing dense genomics, lifestyle, and phenotyping data on \textasciitilde 500,000 volunteers between the ages of 40 – 69 at recruitment, selectively invited to participate in the prospective study based on the NHS register and residing within a traveling distance to one of 22 assessment centers in the United Kingdom (UK). Blood collection for blood cell counts and genomics 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 further 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.

Computation of mtDNA-CN

mtDNA-CN was estimated from DNA derived from buffy coat, using a combination of whole exome and genotyping array data as previously described\textsuperscript{12}. In brief, we first extracted reads from 49,997 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 500k UKB participants\textsuperscript{26}, 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, 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 had cell count outliers.

**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\(^{12}\), extrapolating a nonlinear power-law regression model fit of MAF and effect sizes of the independent common markers (Supplemental Figure S1). The R package genpwr version 1.0.4\(^{28}\) was used for power calculations.

**Single variant analysis**

Single-variant association testing was implemented using the BOLT-LMM mixed-model association testing algorithm\(^{29}\) forcing the 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, then provided a list of 672,036 (85.7%) of the array-based SNPs, restricted to MAF \(\geq 0.001\) and SNP-
missingness ≤ 10%, to allow for construction of the GRM. The merged variants were filtered for variant-
and sample-level missingness > 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 < 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 chrX.

Covariates for analyses included age, sex, and 40 principal components (PCs) capturing genetic ancestry,
and WES batch (45,200 from 50k initial exomes release and 370,222 from the remaining exomes released).
WES batch effects were present due to the use of different oligo lots for exome probe capture in the first
tranche. In addition, genotypes at 129 independent common autosomal variants and four independent
common chrX variants, each previously associated with mtDNA-CN\textsuperscript{12}, 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\textsuperscript{30}. BOLT-LMM internally removed an additional 757 individuals without
all covariates available, yielding an analysis set of 414,665 individuals.

\textit{Aggregate variant analysis}

Aggregate rare variant testing was performed within a mixed-model association testing framework as
implemented in the R Bioconductor package GENESIS\textsuperscript{31} v2.21.4. We restricted the analysis to 24,949,430
(95.9%) variants at ≤ 1% MAF 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 due to genetic relatedness. The
sparse GRM was constructed from kinship estimates provided by the UK Biobank, setting all pairwise
relationships with kinship coefficients < 0.0442 (3rd degree or less) to 0. The previously discovered
common SNPs were not included as covariates for aggregate testing due to computational cost. Next, for each protein coding gene as defined in GENCODE v38 corresponding to GRCh38.p13, rare-variant burden and adjusted, asymptotically independent sequence kernel association test (SKAT) p-values were computed then combined within 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 (df). 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, a SMMAT test was run for nonsynonymous coding sequence variants passing increasing mutational severity thresholds (all nonsynonymous variants, nonsynonymous variants with CADD-PHRED score ≥ 18, predicted loss-of-function (pLoF) variants), and decreasing MAF (≤1%, ≤0.1%, ≤0.01%), such that in total we tested 9 combinations of rare variants assessing various levels of deleteriousness and allele frequency (Supplemental 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 such that singletons and ultra-rare variants with MAC < 6 were included within the aggregate set. This yielded 153,481 total SMMAT tests. We reasoned that since 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 that we conservatively interpreted as outlining four clusters of related p-values (Supplemental 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 / (18557 x 4).

**Gene-level iterative leave-one-out and conditional variant analyses**

For each study-wide significant gene set signal, we used both leave-one-out and gene level conditional analysis 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
a 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 variants with MAC ≥ 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[^4], MsigDb v7.4[^37], and Disgenenet v7[^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, namely the CADD-PHRED ≥ 18 and MAF ≤ 0.001 variant set, was used to obtain z-scores. Since z-scores with extreme values can unduly influence the mean z-score of a gene set, we further tested for robustness of those resulting sets that were significantly enriched at a Bonferroni adjusted family-wise error rate (FWER) of 1.5 x 10^-6 (85 gene sets) by excluding extreme-value SMMAT test z-scores (Supplemental Figure S3), specifically z-scores of TFAM, JAK2, TWNK, and SAMHD1 tests, and re-computing a t-test enrichment p-value.

**PheWAS**
The R package PheWAS[^40] was used to collapse summary diagnoses represented as distinct ICD10 codes in the UKB field 42170 into binary phecodes capturing cases and controls for a variety of disease categories. The analysis was restricted to unrelated individuals using the ‘used.in.pca.calculation’ variable[^36] 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 phecode as a dependent variable regressed onto the rare variant weighted sum score, adjusting for age, sex, and sequencing center. We used the simpleM[^41] method to compute the number of effective tests, which accounted for correlated phecode case and control statuses between clinically related and hierarchical phecodes. 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., \( p \leq 3.0 \times 10^6 \), was used as the Bonferroni multiple test correction denominator for the pheWAS of rare variant carrier status in individual genes. The R package MendelianRandomization\(^{42}\) was used to obtain the inverse variance weighted causal point estimate between mtDNA-CN rare variants in \textit{SAMHD1} and the phecode 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 UK Biobank (UKB)\(^{16}\). mtDNA-CN was estimated from DNA derived from blood buffy coat\(^{12}\).

**Single variants significantly associated with mtDNA-CN**

We performed linear mixed-model association testing\(^{29}\) of single nucleotide variants (SNVs) and small insertion/deletions (indels) with minor allele count (MAC) \( \geq 6 \) in 414,665 UK Biobank exomes of self-reported Whites, adjusting for fixed effects of age, sex, sequencing batch, and random effects due to genetic relatedness and ancestry. We also included genotypes at 129 previously discovered autosomal common variant GWAS loci\(^{12}\) 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 (\( p \leq 7.4 \times 10^{-9} \)) (Table 1A). The most significant SNVs were a pair of rare variants separated by 1.2Mb in \textit{SAMHD1} (NC\_000020.11\:g.36893060C\:T, p.Asp585Asn, \( \beta = +0.61 \) SD [SE = 0.06], \( p\)-value = \( 4.2 \times 10^{-28} \), MAC = 304) and \textit{SPAG4} (NC\_000020.11\:g.35619007C\:G, \( \beta = +0.55 \) SD [SE = 0.06], \( p\)-value = \( 6.4 \times 10^{-22} \), 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 (NC_000020.11:g.1895355G>C, beta = -0.17 SD [SE = 0.02], p-value = 6.6 x 10^{-198}, MAC = 3898), a locus previously identified through the common-variant GWAS\textsuperscript{12}. Additional exome-wide significant single variant associations for mtDNA-CN included a missense variant in JAK2 (NC_000009.12:g.5073770G>T, p.Val617Ile, beta = +0.47 SD [SE = 0.06], p-value = 2.7 x 10^{-17}, MAC = 284), a known dominant somatic gain-of-function mutation associated with myeloproliferative disorders\textsuperscript{43}, and a missense variant within the mtDNA helicase TWNK (NC_000010.11:g.100993257G>A, p.Arg601Gln, beta = -1.07 SD [SE = 0.14], p-value = 1.5 x 10^{-14}, MAC = 48) associated with mitochondrial DNA depletion syndrome 7 (MTDPS7 [MIM: 271245]).

Four variants at study-wide significance were ultra-rare with MAC between six and seven and with concomitantly large effect sizes (mean beta = +2.4 SD, mean SE 0.4 SD). Given the low MAC of these variants, we tested for robustness of these findings by considering sources of false positive effects due to the influence of an outlier mtDNA-CN value, population substructure, or cryptic relatedness. All four variants were robust to the above sources of confounding (Supplemental Figure S4, Supplemental Table S2), and test statistics of variants at MAC 6, 7, and 8 each showed no signs of test statistic inflation (Supplemental Figures S5-6).

**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 similar MAC, we investigated the independence of the variant effects on mtDNA-CN. The missense variants 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^{-21}) and those with only the SAMHD1 variant (40 heterozygotes, p = 0.002) (Supplemental Figure S7). Conditional analysis including both SNPs in the same
model confirmed that the association to mtDNA-CN was driven by the \textit{SAMHD1} variant (conditional p-value $= 1.1 \times 10^{-6}$, \textit{SPAG4} conditional p-value $= 0.59$).

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} solely due to its linkage disequilibrium with \textit{SAMHD1}. Indeed, the 1.2 Mb region between \textit{SPAG4} and \textit{SAMHD1} lies in a previously described recombination desert (Supplemental Figure S8)\textsuperscript{44}, and the surrounding DNA on 20q11 is therefore identical-by-descent in the subset of UK Biobank participants that 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\textsuperscript{31} 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 principal components of ancestry. Due to 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 $\geq 18$, predicted loss-of-function (pLoF) variants), and decreasing MAF ($\leq 1\%$, $\leq 0.1\%$, $\leq 0.01\%$). Eight genes were associated with mtDNA-CN at a study-wide significance cutoff of $p \leq 6.7 \times 10^{-7}$ (Table 1B).

Three additional genes (\textit{POLRMT}, p-value $1.1 \times 10^{-6}$, \textit{CAVIN2}, p-value $1.9 \times 10^{-6}$, and \textit{MGME1}, p-value $2.0 \times 10^{-6}$) had marginal significance (Supplemental Table S3), of which two out of three (\textit{POLRMT}, \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 \textit{TFAM} and \textit{TFB2M} that is required for basal transcription of mitochondrial DNA\textsuperscript{45}. \textit{MGME1} is a mtDNA maintenance exonuclease
and loss-of-function mutations within *MGME1* have been linked with autosomal recessive mtDNA depletion syndrome-11 (MTDPS11 [MIM: 615084])\(^\text{46}\).

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 (Supplemental Table S4), including 4 missense mutations in the most significantly associated gene, *SAMHD1* (gene-based test \(p\)-value = \(4.3 \times 10^{-45}\), Figures 1B-C). Notably, 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\(^\text{47}\) 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, CAVIN2*) had no individual variant that was significantly associated after gene-wide multiple-testing correction, suggesting 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 \(p\)-value = \(1.1 \times 10^{-30}\)) and *TFAM* (SMMAT test \(p\)-value = \(4.3 \times 10^{-15}\)) 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. Notably, only *TWNK* and *TFAM* harbored ultra-rare (MAC < 6) nonsynonymous variants that had significant association with mtDNA-CN. These MAC < 6 variants, coded as a single synthetic allele, were associated with lower mtDNA-CN in *TWNK* (beta = -0.25 SD [SE = 0.04], \(p\)-value = \(2.17 \times 10^{-10}\)) and *TFAM* (beta = -0.37 SD [SE = 0.06], \(p\)-value = \(8.3 \times 10^{-9}\)).

**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 out of 20 genes were nominally associated at a \( p \leq 0.05 \) (3 genes at a Bonferroni-adjusted \( p \)-value \( \leq 0.05 \) accounting for 20 tests), consistent with a highly significant enrichment (exact binomial test \( p \)-value \( 2.9 \times 10^{-6} \), Supplemental 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 Disgenenet\(^{38} \) under the hypothesis that loci harboring rare variants for mtDNA-CN would implicate functionally relevant gene sets related to the genetic architecture of mtDNA-CN. Specifically, 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\(^{39} \). 85 gene sets were significantly associated at a Bonferroni adjusted family-wise error rate of \( 1.5 \times 10^{-6} \) (Supplemental Table S6). To avoid implicating gene sets driven solely by a single gene with a large z-score, we re-computed the enrichment \( p \)-value for the 85 gene sets after excluding extreme value SMMAT test z-scores in each gene set (\( TFAM, JAK2, TFAM, \) and \( SAMHD1 \); Supplemental Figure S3). 19 gene sets retained significance after multiple test correction (\( p \leq 0.05/85 \)), including four that retained Bonferroni-significance accounting for 33,750 tests. Terms central to the synthesis and maintenance of mtDNA were the most significantly enriched, including mtDNA replication (\( p \)-value \( 1.6 \times 10^{-6} \)). mtDNA-CN associated genes were also enriched for medically relevant phenotypes including those related to mitochondrial disease, such as quadriceps muscle weakness (\( p \)-value \( 3.4 \times 10^{-21} \)) and progressive external ophthalmoplegia (\( p \)-value \( 2.5 \times 10^{-16} \)), and bleeding disorders such as macrothrombocytopenia (\( p \)-value \( 7.4 \times 10^{-7} \)). Finally, a gene set composed of common variant implicated loci\(^{12} \) was also significantly enriched for rare variant signal (\( n=111 \) genes in gene set, mean z-score of genes in set = 1.30, background z-score = 0.83, \( p \)-value = \( 3.04 \times 10^{-13} \), \( p \)-value after removing extreme values = \( 1.02 \times 10^{-5} \)).
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 phenome-wide association study (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 < 6 in TFAM that had an effect size < -0.3 SD. PheWAS of the mtDNA-CN rare variant score yielded Bonferroni-significant associations to 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, OR = 822.4 [95% CI = 821.0 – 823.9], t-statistic = 36.9, p-value = 3.2 x 10^-299 for phecode 200, and OR = 166.4 [95% CI = 164.6 – 168.2], t-statistic = 17.1, p-value = 8.5 x 10^-66 for phecode 204.2) and noted that there were no significant associations to 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-value 1.9 x 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. Notably, 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 x 10^{-7}, Supplemental Table S7).

**Discussion**

We conducted an exome-based study of genetic variation associated with mtDNA-CN in 415,422 individuals from the UK Biobank. We report 24 independent signals from 17 loci, including nine SNVs and one rare noncoding variant identified as study-wide independent. Using an iterative leave-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. We note that while the manuscript was in preparation, Chong et al. 2022 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 due to the ~3x larger sample size, this study finds multiple additional loci. Combined with gene-based tests, we also identify 10 loci not detectable by common variant GWAS (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 1.2Mb on 20q11, and formed a single rare, shared haplotype in 262 individuals (haplotype frequency 3.2 x 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 NC_000020.11:g.35619007C>G and NC_000020.11:g.36893060C>T, respectively), suggesting that this haplotype is older than 5-10 generations since the TOPMed cohort is >80% US residents who experienced a bottleneck during the
migration to America\textsuperscript{52}. Indeed, the shared haplotype lies in a known region of very low recombination on 20q\textsuperscript{44}, suggesting 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 \textit{SAMHD1} mutation (NP\_056289.2:p.Asp585Asn), which offers insight into the molecular mechanism behind how variation in \textit{SAMHD1} leads to increased mtDNA-CN. \textit{SAMHD1} encodes a deoxynucleoside triphosphate triphosphohydrolase, a host restriction factor enzyme that depletes dNTP pools and aids in the viral defense response\textsuperscript{53}. The deoxynucleoside products of the enzymatic hydrolysis by SAMHD1 are transported into the mitochondria and salvaged to dNTPs\textsuperscript{54}. While an overabundance of dNTPs could indicate a putative pathway by which heterozygous mutations within \textit{SAMHD1} lead to increased mtDNA-CN, since one determinant of mtDNA-CN is nucleotide metabolism\textsuperscript{12}, we further found that the p.Asp585Asn variant falls within a region of \textit{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 a \textit{SAMHD1}\textsuperscript{4+} AGS5 patient was demonstrated to influence SAMHD1’s alternative role in degrading nascent single-stranded DNA (ssDNA) at stalled replication forks but not influence dNTPase activity\textsuperscript{55}, suggesting that the conditionally independent variant in our study may have an alternative molecular mechanism revolving around failure of ssDNA degradation.

\textit{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\textsuperscript{56-58} and allows for the consideration of the hypothesis that mtDNA-CN may be causally associated with cancer. Notably, conditional analysis of rare variants within \textit{SAMHD1} revealed independent mtDNA-CN-increasing effects at four sites within \textit{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 \textit{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 due to 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, SIRPA) that are pleiotropic for cancer phenotypes. The missense mutation at the JAK2 locus, p.Val617Ile, is a dominant gain-of-function mutation that is causal for myeloproliferative disease$^{43}$, possibly a result of clonal hematopoiesis of intermediate potential (CHIP)$^{59,60}$. Biallelic mutations at TOP3A have been associated with a Bloom-syndrome-like disorder$^{61}$, a rare disorder characterized by predisposition to cancer, albeit with additional mitochondrial dysfunction consistent with the TOP3A function in mtDNA decatenation$^{62}$. CHEK2 is a tumor suppressor gene that encodes a checkpoint protein kinase activated due to 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])$^{63}$. And SIRPA, a gene previously identified in the common variant GWAS for mtDNA-CN$^{12}$, encodes a cell surface receptor for CD47 involved in prevention of autophagocytosis of myeloid cells that is frequently downregulated in a variety of cancers$^{64}$. Nevertheless, while 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 phecodes, 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 (TWNK, TFAM).

The most significant TWNK variant decreasing mtDNA-CN was a missense mutation at NC_000010.11:g.100993257G>A (MAC = 48, beta = -1.09 SD [SE = 0.14], p-value = 1.4 x 10^{-14}), a variant
previously reported in two affected siblings with a diagnosis of autosomal recessive Perrault syndrome-5 (PRLTSS, [MIM: 616138]) who each harbored an additional compound heterozygous mutation in TWNK and whose clinical presentation included muscle weakness and atrophy reminiscent of the broader mitochondrial DNA depletion syndrome-7 (MTDPS7, [MIM: 271245]), the autosomal recessive disorder caused by mutations in TWNK. TFAM encodes a transcription factor part of the initiation complex required for basal transcription of mitochondrial DNA 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, namely at NC_000010.11:g.58388676C>G (MAC = 320, aa residue Q100X, beta = -0.27 SD [SE = 0.058], p-value = 3.0 x 10^-5) and at NC_000010.11:g.58394946A>G (MAC = 81, aa residue K173E, beta = -0.49 SD [SE=0.11], p-value = 1.3 x 10^-5) that 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 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 which could contribute to the missing heritability, including rare variants, an imperfect phenotypic measurement affects the association test power of both common and rare variant GWAS. mtDNA-CN is ‘noisy’ due to confounding from platelet counts and other types of cells, and while we account for these cell counts in our measurement, hidden sources of variability were not considered which 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 $7.4 \times 10^{-9}$, we noted that the present study is underpowered to detect rare single variant associations with mtDNA-CN along the predicted effect sizes for rare variants with MAF < 1% (Supplemental Figure S1). Finally, the study is limited to self-reported White individuals due to a lack of sample size for other genetic ancestry groups.

In conclusion, we have 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, seventeen 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

Whole exome sequencing data for the UKB is available for access through the UKB Research Analysis Platform with an approved UKB application. Summary statistics for mtDNA-CN rare variant association tests are available for download at [https://www.arkinglab.org/resources/](https://www.arkinglab.org/resources/). Scripts and code for generation
of association statistics and additional statistical analyses will be available at
https://github.com/ArkingLab/mtDNA-CN-ExWAS.

Declaration of interests

The authors declare no competing interests.

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 all scripts and code, and wrote and edited the manuscript. W.S., C.S., S.Y., 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.

Web resources

UKB Research Access Platform, https://www.ukbiobank.ac.uk/enable-your-research/research-analysis-platform


MSigDB v7.4, https://www.gsea-msigdb.org/gsea/msigdb/

DisGeNET v7, https://www.disgenet.org

OMIM, https://omim.org
REFERENCES


### Table 1A

Single nucleotide variants associated with mtDNA-CN at exome-wide significance (p ≤ 7.4 x 10⁻⁹).

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Position (bp)a</th>
<th>Reference Allele</th>
<th>Effect Allele</th>
<th>Effect Sizeb</th>
<th>Standard Errorb</th>
<th>Minor Allele Count</th>
<th>P-value</th>
<th>Q-value</th>
<th>Locus / Gene</th>
<th>Functional Effect</th>
<th>Proportion Variance Explainedc</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>36,893,060</td>
<td>C</td>
<td>T</td>
<td>0.61</td>
<td>0.06</td>
<td>304</td>
<td>4.20 x 10⁻²⁸</td>
<td>2.84 x 10⁻²¹</td>
<td>SAMHD1</td>
<td>p.Asp585Asn</td>
<td>0.029%</td>
</tr>
<tr>
<td>20</td>
<td>1,895,355</td>
<td>G</td>
<td>C</td>
<td>-0.17</td>
<td>0.02</td>
<td>3,898</td>
<td>6.60 x 10⁻²⁴</td>
<td>2.23 x 10⁻¹⁷</td>
<td>SIRPA</td>
<td>Intronic variant</td>
<td>0.024%</td>
</tr>
<tr>
<td>20</td>
<td>35,619,007</td>
<td>C</td>
<td>G</td>
<td>0.55</td>
<td>0.06</td>
<td>288</td>
<td>6.40 x 10⁻²²</td>
<td>1.44 x 10⁻¹⁵</td>
<td>SPAG4</td>
<td>Intronic variant</td>
<td>NAd</td>
</tr>
<tr>
<td>9</td>
<td>5,073,770</td>
<td>G</td>
<td>T</td>
<td>0.47</td>
<td>0.06</td>
<td>284</td>
<td>2.70 x 10⁻¹⁵</td>
<td>4.57 x 10⁻¹¹</td>
<td>JAK2</td>
<td>p.Val617Ile</td>
<td>0.017%</td>
</tr>
<tr>
<td>10</td>
<td>100,993,257</td>
<td>G</td>
<td>A</td>
<td>-1.07</td>
<td>0.14</td>
<td>48</td>
<td>1.50 x 10⁻¹⁴</td>
<td>2.03 x 10⁻⁹</td>
<td>TWINK</td>
<td>p.Arg601Gln</td>
<td>0.014%</td>
</tr>
<tr>
<td>17</td>
<td>18,291,048</td>
<td>T</td>
<td>C</td>
<td>0.06</td>
<td>0.01</td>
<td>17,340</td>
<td>6.40 x 10⁻¹⁴</td>
<td>7.22 x 10⁻⁸</td>
<td>TOP3A</td>
<td>Intronic variant</td>
<td>0.014%</td>
</tr>
<tr>
<td>4</td>
<td>184,766,776</td>
<td>G</td>
<td>A</td>
<td>2.36</td>
<td>0.37</td>
<td>7</td>
<td>7.80 x 10⁻¹¹</td>
<td>7.54 x 10⁻⁵</td>
<td>ACSL1</td>
<td>Intronic variant</td>
<td>0.010%</td>
</tr>
<tr>
<td>20</td>
<td>36,898,455</td>
<td>C</td>
<td>G</td>
<td>0.25</td>
<td>0.04</td>
<td>599</td>
<td>3.50 x 10⁻¹⁰</td>
<td>2.96 x 10⁻⁴</td>
<td>SAMHD1</td>
<td>p.Arg531Ser</td>
<td>0.009%</td>
</tr>
<tr>
<td>20</td>
<td>34,280,961</td>
<td>C</td>
<td>T</td>
<td>2.45</td>
<td>0.4</td>
<td>6</td>
<td>5.60 x 10⁻¹⁰</td>
<td>4.21 x 10⁻⁴</td>
<td>AHCY</td>
<td>3' UTR variant</td>
<td>0.009%</td>
</tr>
<tr>
<td>1</td>
<td>205,848,004</td>
<td>A</td>
<td>C</td>
<td>2.36</td>
<td>0.4</td>
<td>6</td>
<td>2.50 x 10⁻⁵</td>
<td>1.69 x 10⁻³</td>
<td>PM20D1</td>
<td>Intronic variant</td>
<td>0.009%</td>
</tr>
<tr>
<td>5</td>
<td>149,332,902</td>
<td>C</td>
<td>A</td>
<td>2.33</td>
<td>0.4</td>
<td>6</td>
<td>4.00 x 10⁻⁵</td>
<td>2.46 x 10⁻³</td>
<td>AFAP1L1</td>
<td>Intronic variant</td>
<td>0.008%</td>
</tr>
</tbody>
</table>

Total

---

*a* Genome coordinates reported in GRCh38  
*b* Effect sizes and standard errors are in units of standard deviation of the scaled mtDNA-CN distribution (mean = 0, std. dev. = 1)  
*c* Proportion of variance (PVE) in mtDNA-CN explained by variant, computed as PVE = Effect-Size² / (Effect-Size² + Std. Error² * Sample-Size) (n=414665)  
*d* Variants present in a shared set of 262 individuals, as part of a rare, shared ancestral haplotype. SPAG4 in linkage disequilibrium with SAMHD1.
Table 1B Aggregate gene-based rare variants associated with mtDNA-CN at study-wide significance

\( p \leq 6.7 \times 10^{-7} \).

<table>
<thead>
<tr>
<th>Number of Qualifying Variant Sites</th>
<th>Number of Non-Reference Alleles</th>
<th>Number of Individuals with Non-Reference Alleles</th>
<th>Cumulative Allele Frequency(^a)</th>
<th>P-value Burden Test</th>
<th>P-value SKAT</th>
<th>P-value SMMAT(^b)</th>
<th>Q-value SMMAT</th>
<th>Gene</th>
<th>Variant Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>258</td>
<td>1,747</td>
<td>1,739</td>
<td>0.0021</td>
<td>4.25 x 10(^{-37})</td>
<td>9.32 x 10(^{-11})</td>
<td>4.27 x 10(^{-45})</td>
<td>6.75 x 10(^{-45})</td>
<td>SAMHD1</td>
<td>Nonsynonymous, CADD-PHRED ≥ 18, MAF ≤ 0.001</td>
</tr>
<tr>
<td>278</td>
<td>1,304</td>
<td>1,296</td>
<td>0.00157</td>
<td>2.20 x 10(^{-18})</td>
<td>6.70 x 10(^{-10})</td>
<td>1.09 x 10(^{-30})</td>
<td>3.46 x 10(^{-26})</td>
<td>TWNK</td>
<td>Nonsynonymous, CADD-PHRED ≥ 18, MAF ≤ 0.0001</td>
</tr>
<tr>
<td>167</td>
<td>1,873</td>
<td>1,870</td>
<td>0.00229</td>
<td>2.74 x 10(^{-10})</td>
<td>4.12 x 10(^{-7})</td>
<td>4.27 x 10(^{-15})</td>
<td>5.62 x 10(^{-11})</td>
<td>TFAM</td>
<td>Nonsynonymous, MAF ≤ 0.01</td>
</tr>
<tr>
<td>636</td>
<td>4,853</td>
<td>4,796</td>
<td>0.00584</td>
<td>2.84 x 10(^{-4})</td>
<td>1.99 x 10(^{-12})</td>
<td>2.04 x 10(^{-14})</td>
<td>2.48 x 10(^{-15})</td>
<td>JAK2</td>
<td>Nonsynonymous, CADD-PHRED ≥ 18, MAF ≤ 0.001</td>
</tr>
<tr>
<td>969</td>
<td>56,902</td>
<td>41,702</td>
<td>0.06849</td>
<td>8.75 x 10(^{-8})</td>
<td>1.56 x 10(^{-4})</td>
<td>3.86 x 10(^{-11})</td>
<td>3.21 x 10(^{-7})</td>
<td>LY75(^c)</td>
<td>Nonsynonymous, MAF ≤ 0.01</td>
</tr>
<tr>
<td>1,137</td>
<td>58,911</td>
<td>43,513</td>
<td>0.07091</td>
<td>2.96 x 10(^{-3})</td>
<td>2.21 x 10(^{-4})</td>
<td>1.75 x 10(^{-10})</td>
<td>1.32 x 10(^{-6})</td>
<td>LY75.CD302(^c)</td>
<td>Nonsynonymous, MAF ≤ 0.01</td>
</tr>
<tr>
<td>309</td>
<td>13,566</td>
<td>13,419</td>
<td>0.01633</td>
<td>2.90 x 10(^{-5})</td>
<td>1.27 x 10(^{-5})</td>
<td>8.41 x 10(^{-9})</td>
<td>5.11 x 10(^{-5})</td>
<td>CLPX</td>
<td>Nonsynonymous, MAF ≤ 0.01</td>
</tr>
<tr>
<td>131</td>
<td>18,663</td>
<td>18,388</td>
<td>0.02408</td>
<td>4.19 x 10(^{-6})</td>
<td>5.33 x 10(^{-4})</td>
<td>4.68 x 10(^{-8})</td>
<td>2.55 x 10(^{-4})</td>
<td>AK2</td>
<td>Nonsynonymous, CADD-PHRED ≥ 18, MAF ≤ 0.01</td>
</tr>
<tr>
<td>457</td>
<td>11,048</td>
<td>10,947</td>
<td>0.0133</td>
<td>2.58 x 10(^{-4})</td>
<td>8.75 x 10(^{-5})</td>
<td>4.19 x 10(^{-7})</td>
<td>2.14 x 10(^{-3})</td>
<td>CHEK2</td>
<td>Nonsynonymous, MAF ≤ 0.01</td>
</tr>
</tbody>
</table>

\(^a\) 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.
**Figures**

**A** Rare, 1.2 Mb shared ancestral haplotype associated with increased mtDNA-CN. A) Rare variants in *SPAG4* (MAC = 288, association with mtDNA-CN $p = 6.4 \times 10^{-22}$, conditional $p = 0.59$) and *SAMHD1* (MAC = 304, association with mtDNA-CN $p = 4.2 \times 10^{-28}$, conditional $p = 1.1 \times 10^{-6}$) were present in a shared set of 262 individuals, representing an ancestral haplotype associated with mtDNA-CN. B) Fine mapping of gene-level rare variant effects via leave-one-out analysis of all 422 protein-coding nonsynonymous variants in *SAMHD1* identifies multiple independent variants associated with mtDNA-CN. Unbroken and dashed red circles indicate gene-level Bonferroni-significant and FDR $\leq 10\%$ associations with mtDNA-CN, respectively. C) Conditional analysis of all 91 MAC $\geq 6$ alleles and a MAC $< 6$ synthetic allele within a single linear model of mtDNA-CN adjusting for age, sex, ancestry PCs 1-40 and exome sequencing batch. Bold $p$-values highlight four Bonferroni-significant conditionally independent associations, including the
missense variant on the rare, shared haplotype (red star). MAC = minor allele count, FDR = false discovery rate, Met = Methionine start codon, Cen = centromere, Tel = telomere.

**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 inverse-variance weighted estimate = 1.29, SE = 0.253, p = $3.1 \times 10^{-7}$.
association betas for disease phecode 174, representing breast cancer risk (n=225,098 females, k=17,517 cases; Supplemental Table S7). The slope of blue line illustrates the inverse-variance weighted causal estimate, bars around points represent 95% confidence intervals.