HGG Advances

ARTICLE

The genetic and phenotypic correlates of mtDNA copy number in a multi-ancestry cohort

Arslan A. Zaidi,[1,5,*] Anurag Verma,[2] Colleen Morse,[2] Penn Medicine BioBank,[3,6] Marylyn D. Ritchie,[1,4] and Iain Mathieson[1,7,*]

Summary

Mitochondrial DNA copy number (mtCN) is often treated as a proxy for mitochondrial (dys-) function and disease risk. Pathological changes in mtCN are common symptoms of rare mitochondrial disorders, but reported associations between mtCN and common diseases vary across studies. To understand the biology of mtCN, we carried out genome- and phenome-wide association studies of mtCN in 30,666 individuals from the Penn Medicine BioBank (PMBB)—a diverse cohort of largely African and European ancestry. We estimated mtCN in peripheral blood using exome sequence data, taking cell composition into account. We replicated known genetic associations of mtCN in the PMBB and found that their effects are highly correlated between individuals of European and African ancestry. However, the heritability of mtCN was much higher among individuals of largely African ancestry ($h^2 = 0.3$) compared with European ancestry individuals($h^2 = 0.1$). Admixture mapping suggests that there are undiscovered variants underlying mtCN that are differentiated in frequency between individuals with African and European ancestry. We show that mtCN is associated with many health-related phenotypes. We discovered robust associations between mtDNA copy number and diseases of metabolically active tissues, such as cardiovascular disease and liver damage, that were consistent across African and European ancestry individuals. Other associations, such as epilepsy and prostate cancer, were only discovered in either individuals with European or African ancestry but not both. We show that mtCN-phenotype associations can be sensitive to blood cell composition and environmental modifiers, explaining why such associations are inconsistent across studies. Thus, mtCN-phenotype associations must be interpreted with care.

Introduction

Mitochondria are vital to cellular function, playing important roles in energy production, calcium signaling, cellular homeostasis, apoptosis, and synthesis of biomolecules. Mitochondrial function is mediated by more than 1,000 proteins—of which only 13 are encoded by the mitochondrial DNA (mtDNA), with the rest encoded by the nuclear genome.[1] Loss of function mutations in these genes can lead to mitochondrial dysfunction, which typically affects multiple systems and tends to be clinically heterogeneous.[2] Considerable effort has been made to understand the genetics of mitochondrial dysfunction through family-based studies of rare mitochondrial diseases.[3] However, the extent to which mitochondrial dysfunction contributes to, or is affected by, common diseases is not well understood.

Practical challenges drive this lack of understanding. Mitochondrial function is difficult to assay in a high-throughput manner. Therefore, most studies use cellular mtDNA content, which can be estimated from sequence data, as a proxy for mitochondrial function. While mtDNA content can be correlated with the respiratory activity of a cell and mtDNA gene expression,[4–7] the relationship is not necessarily linear, and cells may retain as low as 20%–40% of their baseline mtDNA content without a loss in respiratory capacity (see Picard[8] for a review). Both a reduction or elevation of mtDNA copy number can be associated with disease risk.[8,9] However, such associations are inconsistent across studies (e.g., see Filograna et al.[9] for a review), which might be due to lack of power and differences in tissue type used to estimate mtDNA copy number. Hägg et al. and Longchamps et al. are the only well-powered phenome-wide association studies (PheWASs) of mtDNA copy number.[10,11] However, because these studies were both performed in the UK Biobank, it is not clear whether or not phenotypic associations of mtDNA content can be generalized to more diverse cohorts.

In this study, we analyzed genetic and electronic health record data from the Penn Medicine BioBank (PMBB), a large, diverse cohort of African and European ancestry to study the extent to which we can understand the biology underlying genetic and phenotypic correlates of mtDNA copy number. We carried out genome-wide and phenome-wide association studies (GWASs and PheWASs, respectively) separately in two sub-cohorts with largely African (N = 8,598) and European (N = 22,068) ancestry. This allowed us not only to replicate our findings but to

[1]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; [2]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; [3]Center for Translational Bioinformatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; [4]Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[5]Present address: Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, MN, USA
[6]A full list of consortium members is provided in the supplemental information
[7]Lead contact
*Correspondence: aazaidi@umn.edu (A.A.Z.), mathi@pennmedicine.upenn.edu (I.M.)
https://doi.org/10.1016/j.xhgg.2023.100202.

compare and contrast the genetic basis and phenotypic correlates of mtDNA copy number as a function of ancestry.

Material and methods

Description of the dataset

All individuals were patients of the University of Pennsylvania Health System and were enrolled in the Penn Medicine BioBank. Written consent was obtained to collect and store biological specimens and electronic health record (EHR) data and carry out DNA extraction and sequencing. Access and analysis of data for this study were approved by the Institutional Review Board at the University of Pennsylvania.

We started with genetic and EHR data from a total of 39,185 unrelated individuals, who were analyzed in two groups: 10,183 individuals with mixed African and European ancestry and 29,002 individuals with European ancestry (defined broadly), hereafter called AFR and EUR cohorts. Laboratory measurements and disease outcomes were derived from patients' EHR data. Disease outcomes were obtained as ICD-9 and ICD-10 codes, which we mapped to phecodes.[12,13] We defined case/control status for each phecode based on the number of hospital visits, classifying an individual as "case" if they presented in the system with the same phecode at least twice and as "controls" if they were not listed with that phecode at all. Individuals listed once were set to missing. We restricted the analysis to phecodes with more than 20 cases in each of the two cohorts, leading to a total of 1,157 phecodes in the AFR cohort and 1,353 phecodes in the EUR cohort. We also analyzed 25 quantitative laboratory measurements, using the median for each individual if they had multiple measurements. A complete list of phenotypes analyzed for each analysis is available in Tables S1 and S4. For lab measurements, we removed outliers that were >7 standard deviations away from the mean and transformed the values using the optimal Box-Cox power transformation. We used the *boxcox* function in the MASS package[14] in R[15] and estimated λ for the residuals of the following: y ~ poly(Age, 2) + Sex + 20 PCs where y is the lab measurement of interest. Of the initial sample, we had non-missing complete blood count data for 30,666 individuals (EUR = 22,068 and AFR = 8,598) who were then retained for all further analyses.

Calling mtDNA copy number

Mitochondrial DNA copy number represents the number of copies of mtDNA per cell and can be estimated from whole-genome sequence data as twice the ratio of mtDNA depth and autosomal depth. Because exome sequencing involves enrichment of coding sequences in the nuclear genome, we cannot estimate mtDNA copy number in absolute terms (i.e., in number of copies per cell) but can still capture the relative variation in copy number among individuals. To do this, we used *bcftools mpileup* (version 1.12)[16] to call genotypes from reads aligning to the revised Cambridge Reference Sequence (rCRS) of the human mitochondrial genome, filtering out reads with map quality less than 20 (*–m 20*) and base pair quality less than 30 (*–q 30*). Next, we extracted the depth at each position using *bcftools query -f* '%POS% [:DP]', giving us an overall mean depth of 2.8x per site per individual. We observed a spike in sequencing depth between 2.5 and 3 kbp on the rCRS (Figure S1), which has been reported previously.[17] We masked out this region when calculating mean mtDNA depth. We calculated mean autosomal depth across 16,569 sites sampled uniformly at random across the exome. Finally, we took the ratio of mean mtDNA sequencing depth and mean autosomal sequencing depth to get relative mtDNA copy number (rmtCN).

We modeled the log of rmtCN as a function of sex, age, and blood composition using the following linear model in the total sample (AFR and EUR combined) in R[15]:

lrmtCN ~ (Sex + poly(Age, 2)) x (poly(Neutrophil, 2) + poly (Platelets, 2) + poly(Lymphocytes, 2) + poly(Basophils, 2) + poly (Monocytes, 2) + poly(Eosinophils, 2))

This model accounts for nonlinear effects of blood cell counts and allows these effects to vary between males and females and with age. We used the residuals from this model as estimates of mtDNA copy number in all subsequent analyses and referred to them as rlrmtCN.

mtDNA haplogroup calling and ancestry estimation

We used genotypes at 779 mtDNA SNPs that were genotyped on the Illumina Infinium Global Screening Array (GSA) to call haplogroups for each individual with Haplogrep v2.40[18] using the *classify* function with the *–chip* flag. We validated haplogroup calls by calling haplogroups from exome sequence data using off-target reads aligning to the mitochondrial genome. We show that called haplogroups are highly concordant between exome sequence and SNP array data (99% concordance at the top level).

To carry out local ancestry inference, we first phased the genotype data (545,267 SNPs) from the AFR cohort using Beagle version 5.4[19] and then used RFMix[20] to infer local ancestry (k = 2) with genotypes from the 1000 Genomes Project (CEU and YRI)[21] as a reference. We masked out the major histocompatibility locus from chromosome 6 because of the challenge associated with phasing genotypes in this region. We averaged local ancestry for each individual across SNPs that were called with a posterior probability greater than 0.9 to calculate the overall proportion of African and European ancestry. Global ancestry calculated using RFMix was highly correlated ($r^2 = 0.99$) with unsupervised ancestry estimates generated using ADMIXTURE (k = 2).[22]

GWAS, heritability, and PheWAS of mtDNA copy number

We carried out GWAS on rlrmtCN against 10,868,495 autosomal markers, which were imputed using the Michigan Imputation Server[23], with the first 20 genetic principal components (PCs) as covariates. PCs were computed separately within each (AFR and EUR) cohort from a genetic relationship matrix generated using GCTA version 1.93.2beta[24] from common (MAF > 1%), linkage disequilibrium(LD)-pruned (*plink –indep-pairwise 100 10 0.1*[25]) autosomal SNPs that were directly typed on the array.

We carried out admixture mapping in the AFR cohort by testing the association of rlrmtCN with local ancestry at each variant across the genome using a linear model with the global ancestry proportion and genotype for the Duffy-null allele as covariates. The multiple testing burden in admixture mapping tends to be less than that of GWASs because of long-distance correlations in local ancestry that arise due to admixture. We empirically estimated this testing burden using the approach of Shriner et al.[26] Briefly, we estimated the effective number of tests (N_{eff}) by fitting an autoregressive model to the vector of local ancestry for each chromosome of each individual. This was done using the *effectiveSize()* function in the CODA package in R.[15,27] We summed this

number across chromosomes for each individual and then took the mean across individuals to get N_{eff}, which was 17,821 in our case, resulting in a genome-wide significance threshold of $\frac{0.05}{N_{eff}} = 2.81 \times 10^{-6}$.

We used GCTA to estimate the SNP-based heritability of rlrmtCN with the first 20 PCs as fixed effects. We included sex, age and age^2 in addition to the PCs when estimating h_g^2 for lab measurements. We included additional covariates (e.g., Duffy-null genotype, see results section) to determine the source of rlrmtCN heritability in the AFR cohort. The Duffy-null genotype was coded as two variables representing additive ($\in (0, 1, 2)$) and dominant effects (0 for homozygotes and 1 for heterozygotes). To determine if the rlrmtCN heritability in the AFR cohort was driven by unknown differentiated alleles, we selected 21 independent loci from the admixture mapping by clumping at a p value threshold of 0.05 and physical distance of 1 Mb and included the genotypes at these loci as fixed effects (in addition to 20 PCs).

PheWAS for rlrmtCN was carried out using a linear model (for quantitative traits) and logistic regression (for binary traits) with age, age^2, sex, and genetic PCs 1–20. We restricted the analysis to phecodes with at least 20 cases. For lab measurements, we used the trimmed and Box-Cox transformed values described above.

Polygenic risk scores

We constructed polygenic risk scores (PRSs) for 15 blood traits using the variants discovered in Chen et al.[28] We used the summary statistics from the GWASs carried out in individuals of European ancestry (N ≈ 500,000) available from Table S3 of Chen et al.[28] We retained only SNPs for the alleles that matched between Chen et al. and the imputed PMBB genotype data. A comparison between the effect size estimates between Chen et al. and this study is provided in Figure S2. The effect size of one variant (chr1:209451397:G:A) on platelet count as estimated in Chen et al. ($\beta_{Gallele} = -4.4$) was much larger than the other variants and in comparison to its estimate in the PMBB (Figure S2). Their estimate is likely inflated, especially given that the allele is very rare (MAF = < 0.001 in their study). We removed this and another rare variant (chr10:122775741:A:G) that had a large effect on mean corpuscular volume of ($\beta_{Aallele} = -3.49$) before calculating PRSs. This led to a total of 4,394 variants across all traits, which were used to calculate PRSs with the *–score* flag in PLINK.[25] To validate our calculation, we showed that the PRSs were correlated with actual values for traits that were available in the PMBB (i.e., neutrophil, monocyte, platelet, lymphocyte, basophil, eosinophil, and white blood cell count) (Figure S3).

Power to replicate known associations

To estimate the power to replicate known associations for mtDNA copy number, we downloaded Table S6 from Longchamps et al.[11], which contains the list of 129 genome-wide significant SNPs, their positions, and effect sizes. We calculated the power to discover the 110 SNPs that were imputed in the PMBB at the $\alpha = 4.5 \times 10^{-4}$ level of significance (0.05/110 SNPs):

$$\begin{aligned} \mathrm{SE} &= \frac{\sigma}{\sqrt{2nf(1-f)}} \\ \lambda &= \left(\frac{\beta_{gwas}}{\mathrm{SE}}\right)^2 \\ \mathrm{power} &= F^{-1}(\alpha) \end{aligned} \quad , \qquad \text{(Equation 1)}$$

where $\sigma \approx 0.8$ is the residual standard deviation in each cohort (AFR and EUR) after accounting for variance due to age, age^2, sex, and blood cell counts, f is the frequency of the effect allele in the cohort, β_{gwas} is the effect size from the discovery GWAS[11], and F is the cumulative distribution function of a chi-square distribution with non-centrality parameter λ and 1 degree of freedom.

We calculated the heritability explained by GWAS variants separately in each cohort c as $h_c^2 = 2\sum_{i=1}^{m} \widehat{\beta}_{i,c}^2 f_{i,c}(1 - f_{i,c})$ where $\widehat{\beta}_{i,c}$ is the effect size estimate of the i^{th} variant, and $f_{i,c}$ is the minor allele frequency in cohort c.

Analysis of mito-nuclear incompatibility

For the analysis of mito-nuclear incompatibility, we analyzed data from the admixed AFR cohort. We classified haplogroups H, I, J, K, N, R, T, U, V, W, X as "European" and the L haplogroups as "African." Individuals carrying any other haplogroups (N = 271) were removed, resulting in a total of 8,311 individuals. We fit a logistic regression model (linear if the trait was quantitative) with nuclear ancestry, mtDNA haplogroup, and the interaction between the two as predictors and sex, age and age^2 as covariates. We treated mtDNA haplogroup as a factor with the African haplogroup as the reference level.

We calculated power to test for mito-nuclear incompatibility using simulations. We simulated a quantitative trait with effects of sex, age, age^2, nuclear ancestry, mtDNA haplogroup and the interaction between haplogroup and ancestry. We used the effects of sex, age, and age^2 estimated from our data and assumed that the effect of ancestry ranges from 0.05 to 1 (in units of standard deviation of the phenotype). We further assumed a simple model of mito-nuclear incompatibility such that the direction of effect of ancestry is reversed between the two mtDNA haplogroups. We added random noise from a normal distribution with mean zero and standard deviation σ, which was also estimated from the data (after removing variation due to covariates) for each trait separately.

For binary (disease status) traits, we selected the effect size of ancestry ranging from an odds ratio of 1.5–4. Unlike linear models, the power of the test in a logistic regression depends on the intercept term, which specifies the prevalence of the disease in the population. To model this, we fit a logistic regression model to case status for each binary trait with sex, age, and age^2 as predictors. Then, we used the estimated coefficients and the mean value of these predictors from the data to generate the intercept: $\beta_0 = \beta_{intercept} + \beta_{sex} sex + \beta_{age} age + \beta_{age^2} age^2$. Now let $x_j \in \{-1, 1\}$ be an indicator variable coding for the mtDNA haplogroup of individual j, $z_j \in [0, 1]$ be nuclear ancestry, and β_1 be the (assumed) effect size of ancestry. Then, we can simulate case/control status (y_j) for the individual as a bernoulli random variable with probability π_j, where:

$$\pi_j = \frac{exp\{\beta_0 + \beta_1 x_j z_j\}}{1 + exp\{\beta_0 + \beta_1 x_j z_j\}} \qquad \text{(Equation 2)}$$

We fitted a logistic regression model (linear regression for quantitative traits) to the simulated data and evaluated significance of the interaction between mitochondrial and nuclear ancestry if the p value was less than 3.5×10^{-5} (0.05/1137 traits). We repeated this 1,000 times and calculated power as the fraction of iterations where the interaction term was significant.

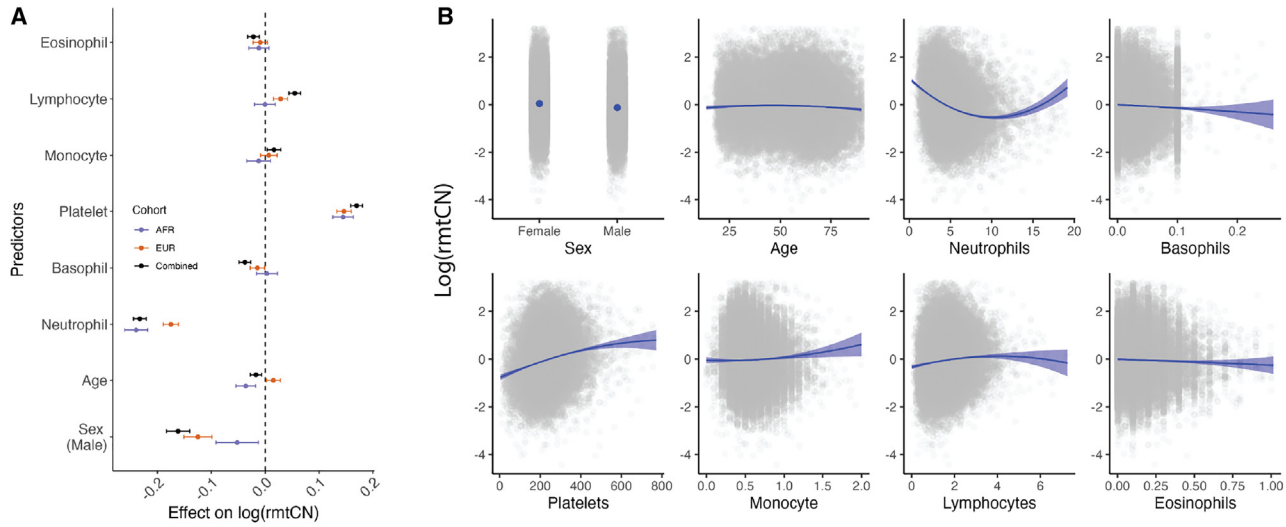

Figure 1. Effect size of sex, age, and blood counts on relative mtDNA copy number (rmtCN)
(A) Effect sizes were estimated jointly using linear regression in the combined sample (AFR+EUR) or separately in the AFR and EUR cohorts. The whiskers represent the 95% confidence intervals of the point estimate. Effect sizes are displayed in units of standard deviation of lrmtCN.
(B) Blue curves represent the predicted values of lrmtCN based on the conditional effects of each predictor (x axis). Actual data are overlaid as gray points. Age is expressed in years, whereas blood counts are expressed in 1,000 cells/ μL of blood.

Results

Peripheral blood mtDNA copy number is a function of cell composition

We estimated mtDNA copy number using the exome sequence data (derived from whole blood) of participants from the PMBB, which has so far recruited more than 175,000 patients with electronic consent through the University of Pennsylvania Health System. We analyzed data from 30,666 unrelated individuals, 22,068 individuals with European ancestry (broadly defined) and 8,598 individuals with mixed African and European ancestry, which we refer to as AFR and EUR cohorts, respectively (analyzed separately). We used the ratio of the mean sequencing depth of off-target reads mapping to the mtDNA to that of reads mapping to an equal number (16,569) of randomly sampled autosomal positions to estimate the average number of mtDNA per cell in whole blood. Note that because exome sequence data are enriched for autosomal reads relative to mtDNA reads, our estimate does not represent the absolute mtDNA content per cell. Instead, we and other studies that rely on exome sequence or array data capture the *relative* number of mtDNA copies per cell (which we refer to as rmtCN).

Inter-individual variation in the mtDNA content of whole blood is a function of blood cell composition.[7] We find that the log of rmtCN is strongly associated with neutrophil and platelet counts and, to a lesser extent, with other cell types (Figure 1) in agreement with previous reports.[8,10,29] The effect of cell composition is consistent in direction between the two cohorts, with neutrophil counts having a negative effect and platelets having a positive effect on rmtCN (Figure 1). The effect of neutrophil count was larger in the AFR cohort compared with the EUR cohort (Figure 1), and this difference is not driven by a confounding effect of ancestry, which is associated with neutrophil counts and mtDNA copy number in opposite directions. One possible explanation for this is that neutrophils in the AFR cohort carry fewer mtDNA copies per cell compared with the EUR cohort.

The effect of cell type composition was also nonlinear (Figure 1), and this needs to be appropriately modeled to ensure that downstream analyses are not driven by variation in cell composition. We modeled the log of rmtCN as a function of sex, age, age^2, and linear and quadratic terms for blood counts (neutrophils, basophils, eosinophils, lymphocytes, monocytes, and platelets). We also included interaction terms to allow the effects of blood counts to vary with age and sex (material and methods). The residuals from this model capture variation in the mean number of mtDNA copies per cell independent of blood cell composition, and we hereafter refer to them as rlrmtCN (residual log of rmtCN). Note, however, that the residuals are not informative about whether mtDNA copy number varies across all cell types uniformly or because of a single cell type. To validate that our model appropriately accounts for blood cell composition, we tested whether rlrmtCN was associated with the Duffy-null allele in the AFR cohort. The Duffy-null allele, because it protects red blood cells from infection by *Plasmodium vivax*, is almost fixed in Africa, while being virtually absent elsewhere.[30] The allele is also one of the strongest known associations for neutropenia (low neutrophil count)[31] and thus is expected to be associated with mtDNA copy number if it captures variation in blood cell composition. We confirm this by showing that the Duffy-null (rs2814778) allele is significantly associated

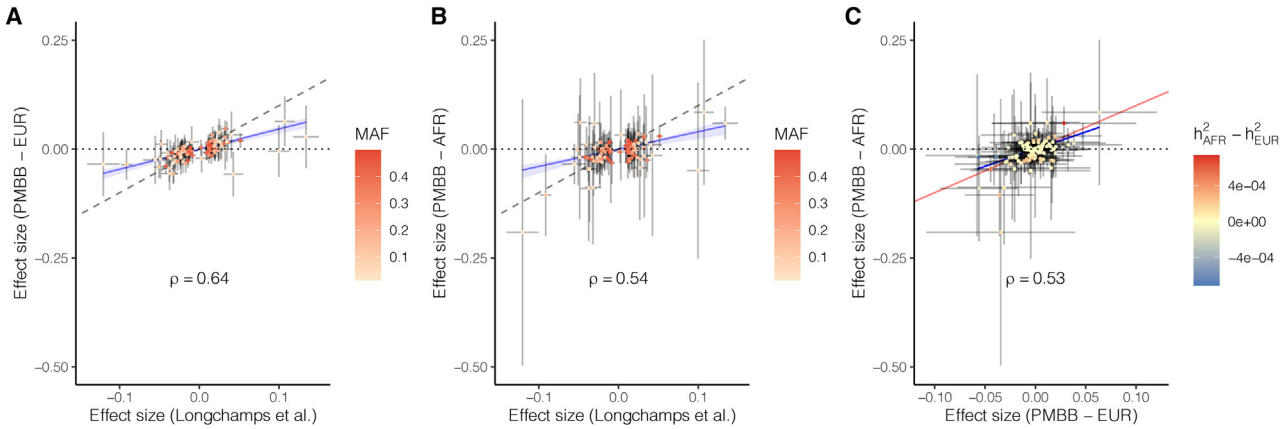

Figure 2. Comparison of effect sizes of mtCN variants discovered in Longchamps et al. and their effects re-estimated in the PMBB(A) EUR cohort and (B) AFR cohort. The PMBB effect sizes were estimated using linear regression with rlrmtCN as response, variant as predictor, and 20 PCs as covariates.(C) Comparison of the effects of the same variants between the AFR and EUR cohorts. The color scale in (A) and (B) represents the minor allele frequency in the original study[11] and in (C) represents the difference between the AFR and EUR cohorts in the variance explained by each locus. The value in each panel represents the correlation coefficient between the two effect sizes.

with the log of rmtCN in the AFR cohort ($\beta_C = -0.3$, $h^2_{explained} = 0.02, p = 1.13 \times 10^{-50}$). In comparison, the effect of the Duffy-null allele on rlrmtCN, i.e., after removing variation due to cell composition, is much smaller $\left(\beta_C = -0.08, h^2_{explained} = 0.002, p = 1.39\times^{-5}\right)$ (Figure S4). Thus, we have largely removed the contribution of neutrophil composition on mtDNA copy number variation. We address any residual association between the Duffy-null allele and mtDNA copy number in a later section.

MtDNA copy number is associated with health-related traits

Because it is correlated with the metabolic activity of the cell, mtDNA copy number is often used as a proxy for mitochondrial (dys-) function[9], and changes in mtDNA copy number are a common symptom and sometimes a cause (e.g., mtDNA depletion syndrome) of mitochondrial diseases.[32] To understand if mtDNA copy number is associated with common diseases, we carried out a PheWAS in the PMBB by testing for associations between rlrmtCN and a range of health-related phenotypes (1,353 in the EUR cohort and 1,157 in the AFR cohort), correcting for sex, age, age^2 and 20 genetic PCs as covariates, separately within each cohort.

MtDNA copy number was associated with many diseases related to metabolically active tissues such as liver, heart, and brain that are common targets of mitochondrial dysfunction.[33–35] For example, in the EUR cohort, rlrmtCN was negatively associated with liver damage (7 phecodes at false discovery rate (FDR) of 0.005 and 9 phecodes at FDR 0.05; e.g., liver abscess, cirrhosis, portal hypertension, and esophageal bleeding, and alcoholism; Figure 2 and Table S1). RlrmtCN was also correlated with aspartate aminotransferase (AST) and total bilirubin in the blood, elevated levels of which are both an indicator of alcohol use and alcohol-related liver damage[36] (Figure 2 and Table S1). While none of these associations were significant at the 0.005 FDR in the AFR cohort, their effects were in the same direction (8/8 phenotypes, binomial p value = 0.008), were correlated ($r_{beta} = 0.65$), and were directionally consistent with previously reported associations with esophageal bleeding and portal hypertension.[10] The association between rlrmtCN and liver damage is attenuated, but it does not disappear, if we include case/control status for alcoholism as a covariate (Table S2). This suggests that the association between rlrmtCN and liver damage may reflect the causal effect of alcohol use on both liver damage and rlrmtCN. This is consistent with an experiment in mice showing that an alcohol binge can lead to drastic changes in mtDNA copy number.[37]

We also observed a positive association between rlrmtCN and cardiac dysfunction—mostly phenotypes related to cardiac dysrhythmias—in the EUR cohort (14 phecodes at FDR 0.005 and 9 phecodes at FDR 0.05; e.g., atrial fibrillation, palpitations, atrial flutter, cardiomyopathy, and mitral valve disease, Figure 2 and Table S1). The associations between rlrmtCN and cardiac phecodes were directionally consistent between the AFR and EUR cohorts (14/14 phecodes, binomial p value = 1.2×10^{-4}) and were correlated ($r_{beta} = 0.56$). The association with cardiac dysrhythmias is also consistent with previous observations of elevated mtDNA copy number in patients with atrial fibrillation.[38,39] However, our associations are in the opposite direction of the negative association with cardiomegaly reported by Hägg et al.[10] and with general cardiovascular disease reported by Ashar et al.[40] We believe that this discrepancy might be explained by differences across studies in how blood cell composition is modeled. Ashar et al.[40] do not fully account for blood cell composition or ancestry, and Hägg et al.[10] only correct for percentage of neutrophils and lymphocytes and total white blood cell count. As an example, we show that the

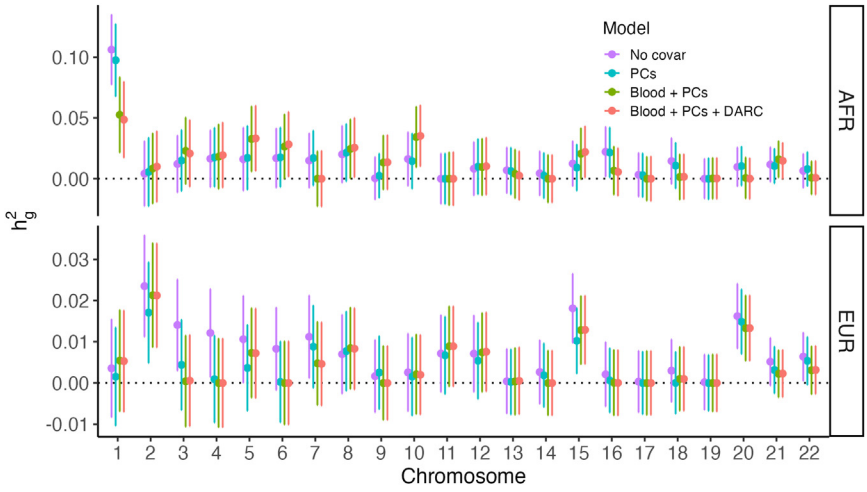

Figure 3. SNP heritability of mtDNA copy number (y axis) contributed by each chromosome (x axis)

The points represent point estimates, and the bars represent the 95% confidence intervals, which were estimated using GCTA.[24] The colors represent different sets of covariates. No covar = no correction for blood composition, sex, age, or PCs; PCs = correction for sex, age, and PCs; PCs + blood = additional correction for blood cell composition; PCs + blood + DARC = additional correction for Duffy-null genotype.

associations between mtDNA copy number and some cardiovascular phenotypes are highly sensitive to how blood cell composition is modeled (Figure S5), suggesting that some previous associations might be driven by blood cell counts as opposed to mtCN per se. The association of rlrmtCN with cardiac dysrhythmia phenotypes (e.g., atrial fibrillation) was less sensitive to blood cell composition as it were positively associated with mtDNA copy number in all models (Figure S5). Altogether, this suggests that mtCN might indeed be a biomarker of some cardiovascular diseases.

The association between some phenotypes and mtDNA copy number was less consistent between the AFR and EUR cohorts. RlrmtCN was negatively associated with epilepsy (3 phecodes at FDR 0.005 and 1 at FDR 0.05), which is a common symptom across mitochondrial disorders, including mtDNA depletion syndrome.[41] But it was not significant in the AFR cohort. In addition to diseases of metabolically active tissues, rlrmtCN was positively associated with prostate cancer and international normalized ratio, which measures the time it takes for blood to clot, and it was negatively associated with rickets in the EUR cohort (0.005 FDR). In the AFR cohort, rlrmtCN was only associated (positively at 0.005 FDR) with iron metabolism disorders.

Ancestry-related differences in the heritability of mtDNA copy number

To study the genetic architecture of mtDNA copy number, we first estimated the SNP heritability (h_g^2) of rlrmtCN in the AFR and EUR cohorts using GCTA[24] with 20 genetic PCs as fixed covariates (material and methods). The h_g^2 of rlrmtCN in the EUR cohort was 0.10 (95% confidence interval [CI]: 0.05–0.14), which overlaps with previous estimates (Hägg et al. = 0.08, Longchamps et al. = 0.07).[10,11] The h_g^2 in the AFR cohort was significantly higher at 0.30 (95% CI: 0.20–0.39), and we explored a number of explanations for this difference.

First, we suspected that the difference in h_g^2 might be driven by known highly differentiated alleles such as the Duffy-null allele, which was not associated with rlrmtCN on a genome-wide level (Figure S4) but may still contribute to rlrmtCN heritability through effects on blood cell composition. This might occur despite corrections for complete blood counts if the counts do not represent the cellular proportions underlying the measured copy number, which in turn could be due to measurement error or because the cell counts were measured at a time or from a sample different from that used for sequencing. This hypothesis was motivated by the observation that neutrophil heritability is also higher in the AFR cohort (Figure S6) and a disproportionately large fraction of rlrmtCN h_g^2 is contributed by chromosome 1, which contains the Duffy locus (Figure 3). However, including the genotype at the Duffy-null allele (rs2814778), which explains most of the heritability in neutrophil counts in the AFR cohort (Figure S7), as a fixed effect in the model does not affect rlrmtCN h_g^2 (Figure 3). Including the genotype for rs73885319 (variant at the *APOL1* locus)—another highly differentiated allele that explains much of the difference in risk of kidney disease between individuals of African and European ancestry[42,43]—also did not change h_g^2 (Figure S8). This suggests that the difference in h_g^2 between the AFR and EUR cohorts cannot be explained by these large effect and highly differentiated alleles.

Second, we asked if rlrmtCN h_g^2 in the AFR cohort could be explained by the heritability underlying blood traits that were not modeled in our analysis. We hypothesized that AST level, which is correlated with rlrmtCN (Figure 2) and also has a higher heritability in the AFR cohort (Figure S6), could be contributing to rlrmtCN h_g^2 in the AFR cohort. However, including AST level as a fixed covariate in our model did not affect h_g^2 estimates (Figure S8), suggesting that this is not the explanation for increased heritability in the AFR cohort.

Third, we investigated whether heritability underlying blood traits that were not measured in the PMBB could be driving h_g^2 in the AFR cohort. To test this, we constructed PRSs for 15 blood traits using effect sizes estimated previously in a GWAS carried out in ≈ 500,000 individuals[28] (material and methods). We validated these scores by showing that the effect sizes of GWAS variants for blood traits that were measured in the PMBB are correlated with their effects in the EUR cohort (Figure S2) and that

the PRS is correlated with the actual phenotype in both cohorts (Figure S3). However, including these PRSs as covariates also did not affect h^2_g estimates (Figure S8). We draw two conclusions from this result: first, that unmeasured blood traits are unlikely to contribute to the difference in rlrmtCN heritability between the AFR and EUR cohorts. Second, the EUR-AFR difference in rlrmtCN h^2_g cannot be explained by measurement noise in blood counts in the PMBB.

Finally, we carried out admixture mapping to identify differentiated alleles that might contribute to rlrmtCN in the AFR cohort. To do this, we tested the association between local ancestry across the genome and rlrmtCN in the AFR cohort with the genome-wide ancestry fraction as a covariate. While we did not discover any loci at the genome-wide level (Figure S9), we show that including the genotypes at the most significant hits (21 independent hits at p value <0.05, material and methods) as covariates in the model substantially reduces rlrmtCN h^2_g in the AFR cohort to 0.15 (95% CI: 0.05–0.14), which overlaps with the h^2_g estimate in the EUR cohort. This suggests that the heritability of rlrmtCN in the AFR cohort might be driven by alleles with large frequency differences between individuals of African and European ancestry. Whether these alleles are associated with mtDNA copy number directly or through their effects on other blood traits will require further investigation.

Similar effects of mtDNA copy number-associated variants in AFR and EUR cohorts

To discover these alleles, we carried out a GWAS of rlrmtCN. We used imputed data and included the first 20 PCs, computed separately in the AFR and EUR cohorts, to correct for population structure. We did not discover any associations at a genome-wide significance threshold of 5×10^{-8} (Figure S10) in either cohort. This includes *TFAM*, which was first identified in a smaller sample of ≈ 10,000 individuals.[44] We then tested if we could replicate other associations discovered previously in much larger GWASs.[10,11,45] These studies were largely based on the same dataset (i.e., the UK Biobank), so we restricted our analysis to variants identified in Longchamps et al., which was the largest study in terms of sample size.[11] Of the 129 independent variants reported in Longchamps et al.[11], 110 were present in our imputed data. Of these, only three were significant at a replication threshold of 4.5×10^{-4} (0.05/110) in the EUR cohort: rs3110823 in the gene *STMP1* ($\beta_{A\ allele} = 0.056,\ p = 2.24 \times 10^{-06}$), rs10419397 near the gene *USHBP1* ($\beta_{A\ allele} = 0.047, p = 8.69\times10^{-7}$), and rs12247015 ($\beta_{A\ allele} = 0.039, p = 1.14\times10^{-5}$) in the 5′ UTR of *TFAM*. This is fewer than the 6.4 associations that we expected to replicate (we had > 80% power to detect 8 loci) based on the effect sizes estimated in Longchamps et al.[11] (Table S3, material and methods). Nevertheless, the effect sizes of the 110 variants were strongly correlated with their effects in the PMBB (Figure 4, $\rho_{EUR} = 0.64, \rho_{AFR} = 0.41$). Note, however, that the PMBB effect sizes are smaller, on average, than the effects reported in Longchamps et al. (Figure 4), which likely explains why we replicated fewer variants than expected. To understand the reason for the downward bias in effect sizes, we considered the possibility that our estimate of mtDNA copy number might be noisier compared with that of Longchamps et al. But, we reject this explanation since the heritability of rlrmtCN in the EUR cohort is similar to previous studies.

The effect sizes of GWAS variants were similar in magnitude and highly correlated between the AFR and EUR cohorts (Figure 4). The GWAS variants also explain a similar fraction of the phenotypic variance in the two cohorts ($h^2_{explained} \sim 0.01$). Thus, the difference in heritability between the two cohorts (see previous section) cannot be explained by a difference in the joint distribution of frequency and effect size at GWAS loci.

No effect of mito-nuclear incompatibility on mtDNA copy number

In a previous study, one of us (A.Z.) found that mtDNA copy number in lymphoblastoid cell lines from admixed individuals was negatively correlated with increasing discordance between the mitochondrial and nuclear genomes such that cells with a higher degree of divergence between nuclear and mitochondrial ancestry exhibited lower mtDNA copy number, on average, than cells where the nuclear and mitochondrial ancestry were similar.[46] This might arise if there was a difference in replication rate between mitochondrial genomes that are more divergent vs. similar in ancestry to the individual's nuclear genome (e.g., due to mito-nuclear incompatibility). We wanted to replicate this result in primary tissue and, thus, analyzed data from a subset of individuals from the AFR cohort with mixed African and European ancestry who carried either a European or African haplogroup (N = 8,311, material and methods). We fitted a linear model with rlrmtCN as the dependent variable and proportion of African ancestry in the nuclear genome, mtDNA ancestry, and the interaction between mtDNA and nuclear ancestry as predictors. The interaction term was not statistically significant ($\beta = -0.31, p = 0.095$; Figure 5) contrary to our expectation under the hypothesis that mito-nuclear ancestry discordance leads to a reduction in mtDNA copy number.[46] The discrepancy between the result shown here and the original study[46] lies in how mito-nuclear discordance is defined. In Zaidi and Makova,[46] mito-nuclear discordance was defined as the total fraction of nuclear ancestry that is different in continental original from the mtDNA. For instance, the discordance of someone with the L mtDNA haplogroup (predominantly found in Africa) and 75%, 25%, and 11% of African, Native American, and European ancestry, respectively, in the nuclear genome would be 0.25 + 0.11 = 0.36. This measure has also been used in other studies to test for mito-nuclear incompatibility in admixed individuals.[47] The problem with this measure, however, is that it captures the main effect of nuclear

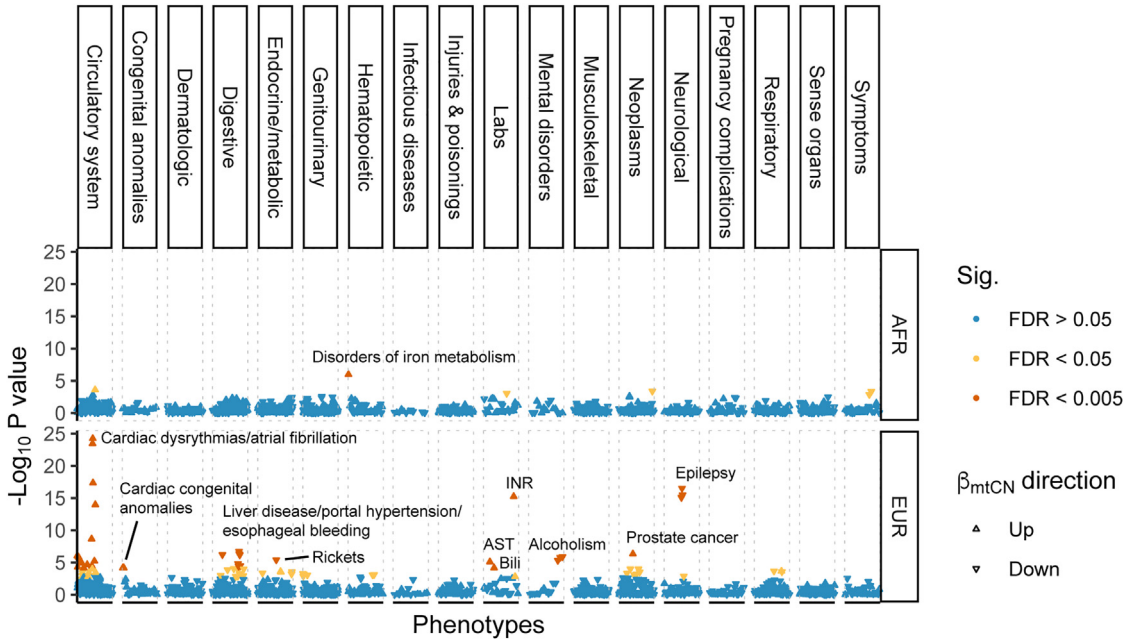

Figure 4. Phenome-wide association study of mtDNA copy numberWe used linear and logistic regression for quantitative and case/control phenotypes, respectively, with sex, age, age^2 and 20 PCs as covariates. Phenotypes are ordered on the x axis, and the y axis shows the –log$_{10}$ p value of the association with mtDNA copy number separately within the AFR and EUR cohorts. Associations that pass the 0.005 and 0.05 false discovery rate are colored in red and yellow, respectively.

ancestry—which is significantly correlated with mtDNA copy number (Figure 5)—if mtDNA haplogroups are non-uniformly distributed in the sample (e.g., 80% African and 20% European in the PMBB). As a result, discordance will be associated with copy number, even if there is no effect of incompatibility. We confirm this by showing that mtDNA haplogroup imbalance in the PMBB also causes mito-nuclear discordance to be negatively associated with rlrmtCN ($\beta = -0.321, p = 3.31\times10^{-6}$). Therefore, the correct way to test for incompatibility is to test for an interaction between nuclear ancestry and mtDNA haplogroup. Reanalysis of data from Zaidi and Makova[46] shows that that the interaction between mtDNA and nuclear ancestry is also not significantly associated with mtDNA copy number in the original study. Altogether, this shows that there is no evidence for an effect of mito-nuclear incompatibility on mtDNA copy number in admixed individuals.

No effect of mito-nuclear incompatibility on health-related traits

We further tested whether there are general phenotypic effects of mito-nuclear incompatibility with a PheWAS on 1,208 health-related phenotypes in the admixed AFR cohort. As before, we fitted ordinary least-squares regression for quantitative traits and logistic regression for binary traits, using nuclear ancestry, mtDNA ancestry, and the interaction between the two as predictors and sex, age, and age^2 as covariates. The proportion of African ancestry in the nuclear genome was positively correlated (at the 0.005 FDR) with a range of phenotypes, including hypertension, hepatitis B, blood pressure (systolic and diastolic), triglyceride levels, serum creatinine levels, and creatine kinase levels, and negatively correlated with neutrophil count (Figure 6)—all consistent with worse health outcomes for people with higher African ancestry and also consistent with previously known associations.[31,48,49] In fact, of the 1,158 phecodes tested (out of total 1,208 phenotypes), 713 were positively associated with African ancestry (binomial test $p = 3.25\times10^{-15}$). In contrast, mtDNA haplogroup was only associated with kidney disease (glomerulonephritis, renal sclerosis) at the 0.05 FDR, with the European haplogroup conferring a higher risk. However, the interaction between mtDNA and nuclear ancestry was not significant at the 0.005 or 0.05 FDR for any phenotype (Figure 6). We show using simulations that, for quantitative traits, we have more than 80% power to detect a negative interaction if the effect of ancestry is larger than 0.5 (in units of standard deviation) (Figure S11). By "negative interaction," we mean that the effect of ancestry on the phenotype is reversed in direction between the two haplogroups but equal in size (material and methods). We have relatively limited power for binary traits but can detect a negative interaction with 80% probability for common diseases (i.e., prevalence >0.35) where the effect of ancestry is greater than an odds ratio of 3.5 (Figure S11). For comparison, the effect of African ancestry on hypertension, which was one of the only binary traits to be significantly associated at the 0.005 FDR, translates to an odds ratio of 3.4. This suggests that the effects of mito-nuclear incompatibility, if present, are not large.

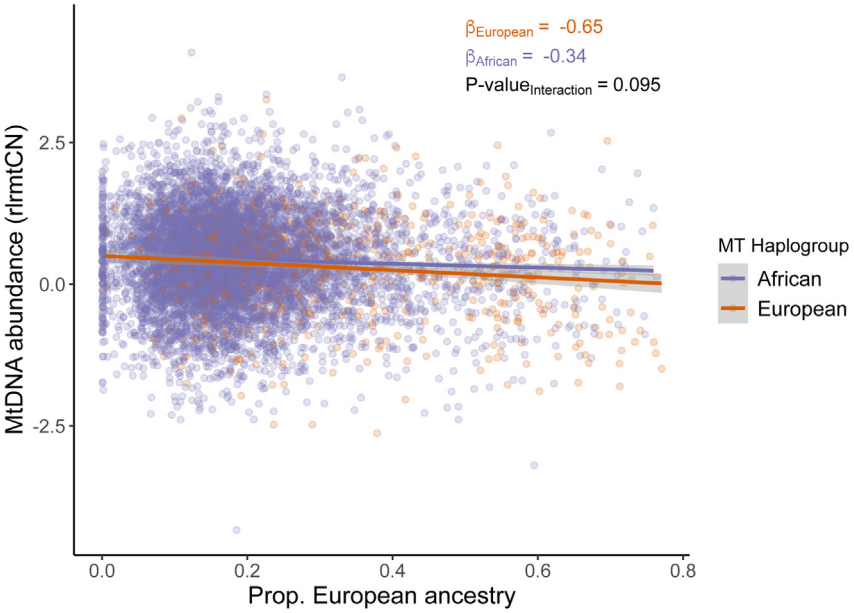

Figure 5. The relationship between nuclear ancestry (x axis) and residual mtDNA copy number (y axis) is similar in different mtDNA backgrounds (colors)
Variation in mtDNA copy number due to sex, age, age^2 and blood counts was removed.

Discussion

The mitochondrial content (total mitochondrial number and volume) of a cell that varies across cell types can be correlated with its mitochondrial activity and bioenergetic needs.[4] Mitochondrial content and activity are also correlated with the number of mtDNA copies in a cell[4], which is easier to assay in a high-throughput manner. As such, there has been interest in using mtDNA copy number as a proxy for mitochondrial (dys-) function in large-scale studies. Many studies have tested for the associations between mtDNA copy number and common diseases (e.g., see Filograna et al.[9] for review) but such associations are inconsistent likely because most studies are under-powered and/or they analyze mtDNA copy number from peripheral blood without appropriately accounting for variation in blood cell composition. This makes it difficult to interpret genetic and phenotypic associations of mtDNA copy number. In this study, we analyzed lab measurements and disease outcomes derived from EHRs as well as genetic data to study the genetics of mtDNA copy number and understand the extent to which it is a useful biomarker of health-related phenotypes in a diverse cohort with African and European ancestry.

MtDNA copy number was associated with several health-related phenotypes, particularly those involving metabolically active tissues such as heart and liver. These associations were largely consistent between the AFR and EUR cohorts. For example, there was a negative association with markers of liver damage and a positive association with phenotypes related to certain cardiovascular diseases. The association between mtDNA copy number and liver damage seems to be mediated largely by alcohol use. The positive association between copy number and atrial fibrillation is consistent with previous studies[38,39] and is thought to be driven by an increase in cell-free circulating mtDNA released by cardiomyocytes in patients with atrial fibrillation, as a result of mitochondrial dysfunction.[39] The association between copy number and other cardiovascular disease (e.g., cardiomegaly) was in the opposite direction to previous reports,[10] and we show that this discrepancy might be due, in part, to differences in how blood cell composition is modeled across studies. Some phenotypic associations were also different within our study between the AFR and EUR cohorts. For example, epilepsy, which is a common symptom of mitochondrial disorders, including mtDNA depletion syndrome, was negatively correlated with mtDNA copy number but only in the EUR cohort. One possibility is that these discrepancies might be driven by differences in environment (e.g., alcohol use) that covary with ancestry.[50,51] Altogether, our results suggest that mtDNA copy number is associated with a range of diseases, but most of these associations are difficult to interpret because they are sensitive to environmental and cellular heterogeneity in peripheral blood and also to methodological choices. As such, phenotypic associations of mtDNA copy number should be interpreted with caution. That said, some associations (e.g., alcohol-related liver disease and atrial fibrillation) were robust and replicated across ancestry groups, suggesting that mtDNA copy number might be a useful biomarker of some diseases.

The genetic architecture of mtDNA copy number was less sensitive to blood cell composition and other modifiers but was strongly associated with ancestry. The heritability of mtDNA copy number was higher in the AFR cohort ($\sim$ 30%) compared with the EUR cohort ($\sim$ 10%). This difference did not appear to be driven by the heritability of blood traits that vary with ancestry (e.g., neutrophil counts and AST levels). In fact, we found that the effect sizes of variants discovered in previous GWASs were highly correlated with their effects in our study, in both EUR and AFR cohorts, despite differences in phenotype construction and ancestry. Interestingly, the difference in heritability of mtDNA copy number between the AFR and EUR cohort was not due to a difference in the frequency or effect sizes of associated variants discovered in previous GWASs.[11] Instead, our admixture mapping analysis suggests that the difference in heritability between the AFR

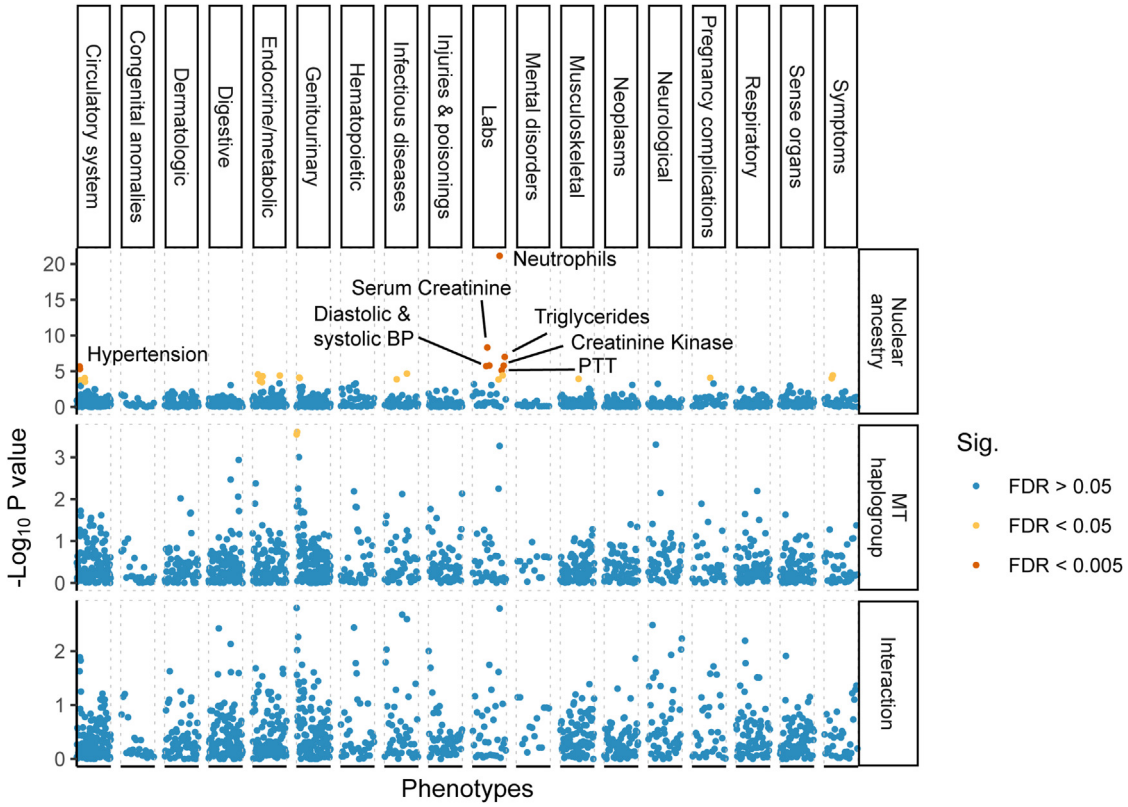

Figure 6. Phenome-wide association of nuclear ancestry (upper), mtDNA haplogrup (middle), and interaction between nuclear and mitochondrial ancestry (lower) in the AFR cohort

The effects were estimated jointly using linear and logistic regression for quantitative and case/control phenotypes, respectively, with sex, age, and age^2 as covariates. Phenotypes are ordered on the x axis grouped into broader categories based on the PheWAS catalog and the y-axis shows -log$_{10}$ p value of association. Association models were either logistic regression (binary phecodes) or linear (lab measurements) with the same covariates: sex, age, and age^2. Associations that pass a false discovery rate of 0.005 are highlighted in red, while those that pass an FDR of 0.05 are shown in yellow.

and EUR cohorts might be driven by variants that are common in AFR but rare in European ancestry populations, and they were therefore not discovered in previous GWASs. If true, larger multi-ethnic studies would be needed to discover these variants and to understand whether they affect mtDNA copy number directly or through other phenotypes that are more heritable among individuals of African ancestry.

We also tested for an effect of mito-nuclear incompatibility on mtDNA copy number. Mito-nuclear incompatibilities have been demonstrated in other organisms (e.g., *Drosophila* and marine copepods[52–54]), but the extent to which they contribute to health risk in humans is widely debated with tangible social consequences, e.g., for mitochondrial replacement therapy.[55–59] An analysis of cell lines from admixed individuals from the 1000 Genomes Project previously showed that increasing mito-nuclear ancestry discordance measured as the fraction of autosomal ancestry that is divergent from mtDNA ancestry leads to a reduction in mtDNA copy number.[46] This was interpreted as an effect of incompatibility between mitochondrial and nuclear ancestry. However, here we show that this does not replicate in primary tissue from a much larger sample of admixed individuals and further show that the original result was due to a statistical artifact that captured the effect of nuclear ancestry, as opposed to that of incompatibility. In conclusion, there is so far no evidence that mito-nuclear incompatibility affects mtDNA copy number in admixed individuals. We also did not detect any significant effects of mito-nuclear incompatibility in a phenome-wide analysis of 1,208 health-related phenotypes. For example, we did not observe an effect of mito-nuclear ancestry interactions on any pregnancy-related phenotypes (e.g., miscarriage, stillbirth, early onset delivery, pre-term birth, and preeclampsia) in contrast to a previous study.[47] That we did not detect such effects in admixed individuals suggests that they do not contribute substantially to variation in medically relevant phenotypes.

Our study highlights the difficulty in interpreting phenotypic associations of mtDNA copy number, as they are mediated by and sensitive to both genetic and environmental modifiers (e.g., ancestry, blood cell composition, and alcohol use). Differences between studies in the distribution of such modifiers and how they are modeled can lead to different results. Blood cell composition can also vary quite drastically with time,[7] and complete blood counts in biobank studies may not come from the same time point as the samples that were used to estimate mtDNA copy number. Thus, even the best methods of correction cannot

guarantee that the associations that we and others observe are independent of the effect of blood counts. This limitation is not unique to mtDNA copy number but to analyses of all cellular readouts (e.g., gene expression) measured in heterogeneous tissues. To further complicate matters, peripheral blood copy number is also a function of cell-free mtDNA, elevated levels of which can be a biomarker of physiological stress and inflammation[60] but which are not measured as part of complete blood counts. These considerations complicate the interpretation of the phenotypic associations of mtDNA copy number. Prospective studies with detailed environmental information and direct quantification of cell-free mtDNA copy number[60], in addition to genetic data and complete blood counts, will be needed to determine whether any associations between copy number and health risk are causal.

Data and code availability

Individual-level genotype and phenotype data from the PMBB are not publicly available due to privacy concerns. However, all summary statistics relevant to this work are made available in Tables S1–S4. The code is publicly available and can be accessed on GitHub (https://github.com/Arslan-Zaidi/mtcn_pmbb). Summary statistics from the GWASs are available on the GWAS catalog under study accessions GWAS Catalog: GCST90267372 and GWAS Catalog: GCST90267373.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2023.100202.

Acknowledgments

We thank the patient-participants of Penn Medicine who consented to participate in this research program. We would also like to thank the Penn Medicine BioBank team and Regeneron Genetics Center for providing genetic variant data for analysis. The PMBB is approved under the University of Pennsylvania IRB protocol #813913. The Penn Medicine BioBank is supported by the Perelman School of Medicine at University of Pennsylvania, a gift from the Smilow family, and the National Center for Advancing Translational Sciences of the National Institutes of Health under CTSA award number UL1TR001878. This study was funded by NIGMS awards K99GM137076 (to A.Z.) and R35GM133708 (to I.M.). The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Received: March 2, 2023
Accepted: April 25, 2023

References

1. Rath, S., Sharma, R., Gupta, R., Ast, T., Chan, C., Durham, T.J., Goodman, R.P., Grabarek, Z., Haas, M.E., Hung, W.H.W., et al. (2021). MitoCarta3.0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations. Nucleic Acids Res. *49*, D1541–D1547.
2. Taylor, R.W., and Turnbull, D.M. (2005). Mitochondrial DNA mutations in human disease. Nat. Rev. Genet. *6*, 389–402.
3. Chinnery, P.F. (2021). Primary Mitochondrial Disorders Overview (GeneReviews®).
4. D'Erchia, A.M., Atlante, A., Gadaleta, G., Pavesi, G., Chiara, M., De Virgilio, C., Manzari, C., Mastropasqua, F., Prazzoli, G.M., Picardi, E., et al. (2015). Tissue-specific mtDNA abundance from exome data and its correlation with mitochondrial transcription, mass and respiratory activity. Mitochondrion *20*, 13–21.
5. Bentlage, H.A., and Attardi, G. (1996). Relationship of genotype to phenotype in fibroblast-derived transmitochondrial cell lines carrying the 3243 mutation associated with the melas encephalomyopathy: shift towards mutant genotype and role of mtDNA copy number. Hum. Mol. Genet. *5*, 197–205.
6. Lee, S.R., Heo, H.J., Jeong, S.H., Kim, H.K., Song, I.S., Ko, K.S., Rhee, B.D., Kim, N., and Han, J. (2015). Low abundance of mitochondrial DNA changes mitochondrial status and renders cells resistant to serum starvation and sodium nitroprusside insult. Cell Biol. Int. *39*, 865–872.
7. Rausser, S., Trumpff, C., McGill, M.A., Junker, A., Wang, W., Ho, S.H., Mitchell, A., Karan, K.R., Monk, C., Segerstrom, S.C., et al. (2021). Mitochondrial phenotypes in purified human immune cell subtypes and cell mixtures. Elife *10*, e70899.
8. Picard, M. (2021). Blood mitochondrial DNA copy number: what are we counting? Mitochondrion *60*, 1–11.
9. Filograna, R., Mennuni, M., Alsina, D., and Larsson, N.G. (2021). Mitochondrial DNA copy number in human disease: the more the better? FEBS Lett. *595*, 976–1002.
10. Hägg, S., Jylhävä, J., Wang, Y., Czene, K., and Grassmann, F. (2021). Deciphering the genetic and epidemiological landscape of mitochondrial DNA abundance. Hum. Genet. *140*, 849–861.
11. Longchamps, R.J., Yang, S.Y., Castellani, C.A., Shi, W., Lane, J., Grove, M.L., Bartz, T.M., Sarnowski, C., Liu, C., Burrows, K., et al. (2022). Genome-wide analysis of mitochondrial DNA copy number reveals loci implicated in nucleotide metabolism, platelet activation, and megakaryocyte proliferation. Hum. Genet. *141*, 127–146.
12. Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D.R., Roden, D.M., and Crawford, D.C. (9 2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics *26*, 1205–1210.
13. Bastarache, L. (2021). Using phecodes for research with the electronic health record: from PheWAS to PheRS. Annu. Rev. Biomed. Data Sci. *4*, 1–19.
14. Venables, W.N., and Ripley, B.D. (2002). *Modern Applied Statistics With S* Fourth (Springer).
15. R Core Team. *R* (2021). A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).
16. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.
17. García-Olivares, V., Muñoz-Barrera, A., et al. (2021). A benchmarking of human mitochondrial DNA haplogroup classifiers from whole-genome and whole-exome sequence data. Sci. Rep. *11*, 1–11.

18. Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.J., Kronenberg, F., Salas, A., and Schönherr, S. (2016). HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. *44*, W58–W63.
19. Browning, B.L., Tian, X., Zhou, Y., and Browning, S.R. (2021). Fast two-stage phasing of large-scale sequence data. Am. J. Hum. Genet. *108*, 1880–1890.
20. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288.
21. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.
22. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.
23. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284–1287.
24. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
25. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7.
26. Shriner, D., Adeyemo, A., and Rotimi, C.N. (2011). Joint ancestry and association testing in admixed individuals. PLoS Comput. Biol. *7*, e1002325.
27. Plummer, M., Best, N., et al. (2006). CODA: convergence diagnosis and output analysis for MCMC. R. News *6*, 7–11.
28. Chen, M.H., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., et al. (2020). Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. Cell *182*, 1198–1213.e14.
29. Knez, J., Winckelmans, E., Plusquin, M., Thijs, L., Cauwenberghs, N., Gu, Y., Staessen, J.A., Nawrot, T.S., and Kuznetsova, T. (2016). Correlates of peripheral blood mitochondrial DNA content in a general population. Am. J. Epidemiol. *183*, 138–146.
30. Howes, R.E., Patil, A.P., Piel, F.B., Nyangiri, O.A., Kabaria, C.W., Gething, P.W., Zimmerman, P.A., Barnadas, C., Beall, C.M., Gebremedhin, A., et al. (2011). The global distribution of the Duffy blood group. Nat. Commun. *2*.
31. Reich, D., Nalls, M.A., Kao, W.H., Akylbekova, E.L., Tandon, A., Patterson, N., Mullikin, J., Hsueh, W.C., Cheng, C.Y., Coresh, J., et al. (2009). Reduced neutrophil count in people of African descent is due to a regulatory variant in the duffy antigen receptor for chemokines gene. PLoS Genet. *5*, e1000360. ed, Visscher, P. M.
32. Ng, Y.S., and Turnbull, D.M. (2016). Mitochondrial disease: genetics and management. J. Neurol. *263*, 179–191.
33. Parikh, S., Goldstein, A., Koenig, M.K., Scaglia, F., Enns, G.M., Saneto, R., Anselm, I., Cohen, B.H., Falk, M.J., Greene, C., et al. (2015). Diagnosis and management of mitochondrial disease: a consensus statement from the Mitochondrial Medicine Society. Genet. Med. *17*, 689–701.
34. Suomalainen, A., and Battersby, B.J. (2018). Mitochondrial diseases: the contribution of organelle stress responses to pathology. Nat. Rev. Mol. Cell Biol. *19*, 77–92.
35. Russell, O.M., Gorman, G.S., Lightowlers, R.N., and Turnbull, D.M. (2020). Mitochondrial diseases: hope for the future. Cell *181*, 168–188.
36. Agarwal, S., Fulgoni, V.L., and Lieberman, H.R. (2016). Assessing alcohol intake & its dosedependent effects on liver enzymes by 24-h recall and questionnaire using NHANES 2001-2010 data. Nutr. J. *15*, 62.
37. Mansouri, A., Gaou, I., De Kerguenec, C., Amsellem, S., Haouzi, D., Berson, A., Moreau, A., Feldmann, G., Lettéron, P., Pessayre, D., and Fromenty, B. (1999). An alcoholic binge causes massive degradation of hepatic mitochondrial DNA in mice. Gastroenterology *117*, 181–190.
38. Zhang, Y., Guallar, E., Ashar, F.N., Longchamps, R.J., Castellani, C.A., Lane, J., Grove, M.L., Coresh, J., Sotoodehnia, N., Ilkhanoff, L., et al. (2017). Association between mitochondrial DNA copy number and sudden cardiac death: findings from the Atherosclerosis Risk in Communities study (ARIC). Eur. Heart J. *38*, 3443–3448.
39. Wiersma, M., van Marion, D.M.S., Bouman, E.J., Li, J., Zhang, D., Ramos, K.S., Lanters, E.A.H., de Groot, N.M.S., and Brundel, B.J.J.M. (2020). Cell-free circulating mitochondrial DNA: a potential blood-based marker for atrial fibrillation. Cells *9*, 1159.
40. Ashar, F.N., Zhang, Y., Longchamps, R.J., Lane, J., Moes, A., Grove, M.L., Mychaleckyj, J.C., Taylor, K.D., Coresh, J., Rotter, J.I., et al. (2017). Association of mitochondrial DNA copy number with cardiovascular disease. JAMA Cardiol. *2*, 1247–1255.
41. El-Hattab, A.W., and Scaglia, F. (2013). Mitochondrial DNA depletion syndromes: review and updates of genetic basis, manifestations, and therapeutic options. Neurotherapeutics *10*, 186–198.
42. Genovese, G., Friedman, D.J., Ross, M.D., Lecordier, L., Uzureau, P., Freedman, B.I., Bowden, D.W., Langefeld, C.D., Oleksyk, T.K., Uscinski Knob, A.L., et al. (2010). Association of trypanolytic ApoL1 variants with kidney disease in African Americans. Science *329*, 841–845.
43. Limou, S., Nelson, G.W., Kopp, J.B., and Winkler, C.A. (2014). APOL1 kidney risk alleles: population genetics and disease associations. Adv. Chronic Kidney Dis. *21*, 426–433.
44. Cai, N., Li, Y., Chang, S., Liang, J., Lin, C., Zhang, X., Liang, L., Hu, J., Chan, W., Kendler, K.S., et al. (2015). Genetic control over mtDNA and its relationship to major depressive disorder. Curr. Biol. *25*, 3170–3177.
45. 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.
46. Zaidi, A.A., and Makova, K.D. (2019). Investigating mitonuclear interactions in human admixed populations. Nat. Ecol. Evol. *3*, 213–222.
47. Crawford, N., Prendergast, D., Oehlert, J.W., Shaw, G.M., Stevenson, D.K., Rappaport, N., Sirota, M., Tishkoff, S.A., and Sondheimer, N. (2018). Divergent patterns of mitochondrial and nuclear ancestry are associated with the risk for preterm birth. J. Pediatr. *194*, 40–46.e4.
48. Udler, M.S., Nadkarni, G.N., Belbin, G., Lotay, V., Wyatt, C., Gottesman, O., Bottinger, E.P., Kenny, E.E., and Peter, I.

(2015). Effect of genetic African ancestry on EGFR and kidney disease. J. Am. Soc. Nephrol. *26*, 1682–1692.
49. Zilbermint, M., Hannah-Shmouni, F., and Stratakis, C.A. (2019). Genetics of hypertension in african Americans and others of African descent. Int. J. Mol. Sci. *20*, 1081.
50. Samokhvalov, A.V., Irving, H., Mohapatra, S., and Rehm, J. (2010). Alcohol consumption, unprovoked seizures, and epilepsy: a systematic review and meta-analysis. Epilepsia *51*, 1177–1184.
51. Wu, L.T., Woody, G.E., Yang, C., Pan, J.J., and Blazer, D.G. (2011). Racial/ethnic variations in substance-related disorders among adolescents in the United States. Arch. Gen. Psychiatry *68*, 1176–1185.
52. Holmbeck, M.A., Donner, J.R., Villa-Cuesta, E., and Rand, D.M. (2015). A Drosophila model for mito-nuclear diseases generated by an incompatible interaction between tRNA and tRNA synthetase. Dis. Model. Mech. *8*, 843–854.
53. Burton, R.S., Ellison, C.K., and Harrison, J.S. (2006). The sorry state of F2 hybrids: consequences of rapid mitochondrial DNA evolution in allopatric populations. Am. Nat. *168*, S14–S24.
54. Meiklejohn, C.D., Holmbeck, M.A., Siddiq, M.A., Abt, D.N., Rand, D.M., and Montooth, K.L. (2013). An incompatibility between a mitochondrial tRNA and its nuclear-encoded tRNA synthetase compromises development and fitness in Drosophila. B.A. Payseur, ed. *9*, e1003238.
55. Reinhardt, K., Dowling, D.K., and Morrow, E.H. (2013). Medicine. Mitochondrial replacement, evolution, and the clinic. Science *341*, 1345–1346.
56. Chou, J.-Y., and Leu, J.-Y. (2015). The Red Queen in mitochondria: cyto-nuclear co-evolution, hybrid breakdown and human disease. Front. Genet. *6*, 187.
57. Morrow, E.H., Reinhardt, K., Wolff, J.N., and Dowling, D.K. (2015). Risks inherent to mitochondrial replacement. EMBO Rep. *16*, 541–544.
58. Claiborne, A.B., English, R.A., and Kahn, J.P. (2016). Finding an ethical path forward for mitochondrial replacement. Science *351*, 668–670.
59. Eyre-Walker, A. (2017). Mitochondrial replacement therapy: are mito-nuclear interactions likely to Be a problem? Genetics *205*, 1365–1372.
60. Trumpff, C., Michelson, J., Lagranha, C.J., Taleon, V., Karan, K.R., Sturm, G., Lindqvist, D., Fernström, J., Moser, D., Kaufman, B.A., and Picard, M. (2021). Stress and circulating cell-free mitochondrial DNA: a systematic review of human studies, physiological considerations, and technical recommendations. Mitochondrion *59*, 225–245.

HGGA, Volume *4*

Supplemental information

The genetic and phenotypic correlates of mtDNA copy number in a multi-ancestry cohort

Arslan A. Zaidi, Anurag Verma, Colleen Morse, Penn Medicine BioBank, Marylyn D. Ritchie, and Iain Mathieson

References

1. Rath, S., Sharma, R., *et al.* MitoCarta3.0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations. *Nucleic Acids Research* **49,** D1541–D1547 (D1 2021).

2. Taylor, R. W. & Turnbull, D. M. Mitochondrial DNA mutations in human disease. *Nature Reviews Genetics* **6,** 389–402 (5 2005).

3. Chinnery, P. F. Primary Mitochondrial Disorders Overview. *GeneReviews®* (2021).

4. D'Erchia, A. M., Atlante, A., *et al.* Tissue-specific mtDNA abundance from exome data and its correlation with mitochondrial transcription, mass and respiratory activity. *Mitochondrion* **20,** 13–21 (2015).

5. Bentlage, H. A. & Attardi, G. Relationship of Genotype to Phenotype in Fibroblast-derived Transmitochondrial Cell Lines Carrying the 3243 Mutation Associated with the Melas Encephalomyopathy: Shift towards Mutant Genotype and Role of mtDNA Copy Number. *Human Molecular Genetics* **5,** 197–205 (2 1996).

6. Lee, S. R., Heo, H. J., *et al.* Low abundance of mitochondrial DNA changes mitochondrial status and renders cells resistant to serum starvation and sodium nitroprusside insult. *Cell Biology International* **39,** 865–872 (7 2015).

7. Rausser, S., Trumpff, C., *et al.* Mitochondrial phenotypes in purified human immune cell subtypes and cell mixtures. *eLife* **10** (2021).

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

9. Filograna, R., Mennuni, M., *et al.* Mitochondrial DNA copy number in human disease: the more the better? *FEBS Letters* **595,** 976–1002 (8 2021).

10. Hägg, S., Jylhävä, J., *et al.* Deciphering the genetic and epidemiological landscape of mitochondrial DNA abundance. *Human Genetics* **140,** 849–861 (2021).

11. Longchamps, R. J., Yang, S. Y., *et al.* Genome-wide analysis of mitochondrial DNA copy number reveals loci implicated in nucleotide metabolism, platelet activation, and megakaryocyte proliferation. *Human Genetics* **141,** 127–146 (2022).

12. Knez, J., Winckelmans, E., *et al.* Correlates of Peripheral Blood Mitochondrial DNA Content in a General Population. *American Journal of Epidemiology* **183,** 138–146 (2015).

13. Howes, R. E., Patil, A. P., *et al.* The global distribution of the Duffy blood group. *Nature Communications 2011 2:1* **2,** 1–10 (1 2011).

14. Reich, D., Nalls, M. A., *et al.* Reduced Neutrophil Count in People of African Descent Is Due To a Regulatory Variant in the Duffy Antigen Receptor for Chemokines Gene. *PLoS Genetics* **5** (ed Visscher, P. M.) e1000360 (2009).

15. Ng, Y. S. & Turnbull, D. M. Mitochondrial disease: genetics and management. *Journal of Neurology* **263,** 179–191 (1 2016).

16. Parikh, S., Goldstein, A., *et al.* Diagnosis and management of mitochondrial disease: A consensus statement from the Mitochondrial Medicine Society. *Genetics in Medicine* **17,** 689–701 (9 2015).

17. Suomalainen, A. & Battersby, B. J. Mitochondrial diseases: The contribution of organelle stress responses to pathology. *Nature Reviews Molecular Cell Biology* **19,** 77–92 (2 2018).

18. Russell, O. M., Gorman, G. S., *et al.* Mitochondrial Diseases: Hope for the Future. *Cell* **181,** 168–188 (1 2020).

19. Agarwal, S., Fulgoni, V. L., *et al.* Assessing alcohol intake & its dosedependent effects on liver enzymes by 24-h recall and questionnaire using NHANES 2001-2010 data. *Nutrition Journal* **15,** 1–12 (1 2016).

20. Mansouri, A., Gaou, I., *et al.* An alcoholic binge causes massive degradation of hepatic mitochondrial DNA in mice. *Gastroenterology* **117,** 181–190 (1 1999).

21. Zhang, Y., Guallar, E., *et al.* Association between mitochondrial DNA copy number and sudden cardiac death: Findings from the Atherosclerosis Risk in Communities study (ARIC). *European Heart Journal* **38,** 3443–3448 (2017).

22. Wiersma, M., van Marion, D. M., *et al.* Cell-Free Circulating Mitochondrial DNA: A Potential Blood-Based Marker for Atrial Fibrillation. *Cells* **9** (5 2020).

23. Ashar, F. N., Zhang, Y., *et al.* Association of mitochondrial DNA copy number with cardiovascular disease. *JAMA Cardiology* **2,** 1247–1255 (2017).

24. El-Hattab, A. W. & Scaglia, F. Mitochondrial DNA Depletion Syndromes: Review and Updates of Genetic Basis, Manifestations, and Therapeutic Options. *Neurotherapeutics* **10,** 186–198 (2 2013).

25. Yang, J., Lee, S. H., *et al.* GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics* **88,** 76–82 (1 2011).

26. Genovese, G., Friedman, D. J., *et al.* Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science* **329,** 841–5 (2010).

27. Limou, S., Nelson, G. W., *et al.* APOL1 kidney risk alleles: population genetics and disease associations. *Adv Chronic Kidney Dis* **21,** 426–33 (2014).

28. Chen, M. H., Raffield, L. M., *et al.* Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell* **182,** 1198–1213.e14 (5 2020).

29. Cai, N., Li, Y., *et al.* Genetic Control over mtDNA and Its Relationship to Major Depressive Disorder. *Current Biology* **25,** 3170–3177 (24 2015).

30. Chong, M., Mohammadi-Shemirani, P., *et al.* GWAS and ExWAS of blood Mitochondrial DNA copy number identifies 71 loci and highlights a potential causal role in dementia. *eLife* **11** (2022).

31. Zaidi, A. A. & Makova, K. D. Investigating mitonuclear interactions in human admixed populations. *Nature Ecology and Evolution* **3,** 213–222 (2019).

32. Crawford, N., Prendergast, D., *et al.* Divergent Patterns of Mitochondrial and Nuclear Ancestry Are Associated with the Risk for Preterm Birth. *The Journal of Pediatrics* **194,** 40–46.e4 (2018).

591 33. Udler, M. S., Nadkarni, G. N., *et al.* Effect of genetic African ancestry on EGFR and kidney disease.
592 *Journal of the American Society of Nephrology* **26,** 1682–1692 (2015).

593 34. Zilbermint, M., Hannah-Shmouni, F., *et al.* Genetics of Hypertension in African Americans and
594 Others of African Descent. *International Journal of Molecular Sciences* **20,** 1081 (2019).

595 35. Samokhvalov, A. V., Irving, H., *et al.* Alcohol consumption, unprovoked seizures, and epilepsy: A
596 systematic review and meta-analysis. *Epilepsia* **51,** 1177–1184 (7 2010).

597 36. Wu, L. T., Woody, G. E., *et al.* Racial/Ethnic Variations in Substance-Related Disorders Among
598 Adolescents in the United States. *Archives of General Psychiatry* **68,** 1176–1185 (11 2011).

599 37. Holmbeck, M. A., Donner, J. R., *et al.* A Drosophila model for mito-nuclear diseases generated
600 by an incompatible interaction between tRNA and tRNA synthetase. *DMM Disease Models and*
601 *Mechanisms* **8,** 843–854 (8 2015).

602 38. Burton, R. S., Ellison, C. K., *et al.* The sorry state of F2 hybrids: consequences of rapid mitochon-
603 drial DNA evolution in allopatric populations. *The American naturalist* **168,** S14–S24 (2006).

604 39. Meiklejohn, C. D., Holmbeck, M. A., *et al.* An Incompatibility between a Mitochondrial tRNA and
605 Its Nuclear-Encoded tRNA Synthetase Compromises Development and Fitness in Drosophila. *PLoS*
606 *Genetics* **9** (ed Payseur, B. A.) e1003238 (1 2013).

607 40. Reinhardt, K., Dowling, D. K., *et al.* Medicine. Mitochondrial replacement, evolution, and the clinic.
608 *Science* **341,** 1345–6 (2013).

609 41. Chou, J.-Y. & Leu, J.-Y. The Red Queen in mitochondria: cyto-nuclear co-evolution, hybrid break-
610 down and human disease. *Frontiers in Genetics* **6,** 187 (2015).

611 42. Morrow, E. H., Reinhardt, K., *et al.* Risks inherent to mitochondrial replacement. *EMBO reports*
612 **16,** 541–544 (5 2015).

613 43. Claiborne, A. B., English, R. A., *et al.* Finding an ethical path forward for mitochondrial replace-
614 ment. *Science* **351,** 668–670 (6274 2016).

615 44. Eyre-Walker, A. Mitochondrial Replacement Therapy: Are Mito-nuclear Interactions Likely To Be
616 a Problem? *Genetics* **205,** 1365–1372 (4 2017).

617 45. Trumpff, C., Michelson, J., *et al.* Stress and circulating cell-free mitochondrial DNA: A systematic
618 review of human studies, physiological considerations, and technical recommendations. *Mitochon-*
619 *drion* **59** (2021).

620 46. Denny, J. C., Ritchie, M. D., *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan
621 to discover gene–disease associations. *Bioinformatics* **26,** 1205–1210 (9 2010).

622 47. Bastarache, L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to
623 PheRS. *Annual review of biomedical data science* **4,** 1–19 (2021).

624 48. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S* Fourth. ISBN 0-387-95457-0
625 (Springer, New York, 2002).

626 49. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Sta-
627 tistical Computing (Vienna, Austria, 2021).

50. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987 (21 2011).

51. García-Olivares, V., Muñoz-Barrera, A., *et al.* A benchmarking of human mitochondrial DNA haplogroup classifiers from whole-genome and whole-exome sequence data. *Scientific Reports* **11,** 1–11 (2021).

52. Weissensteiner, H., Pacher, D., *et al.* HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Research* **44,** W58–W63 (W1 2016).

53. Fast two-stage phasing of large-scale sequence data. *American Journal of Human Genetics* **108,** 1880–1890 (2021).

54. Maples, B. K., Gravel, S., *et al.* RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *American journal of human genetics* **93,** 278–88 (2013).

55. Auton, A., Abecasis, G. R., *et al.* A global reference for human genetic variation. *Nature 2015 526:7571* **526,** 68–74 (7571 2015).

56. Alexander, D. H., Novembre, J., *et al.* Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19,** 1655–64 (2009).

57. Das, S., Forer, L., *et al.* Next-generation genotype imputation service and methods. *Nature Genetics* **48,** 1284–1287 (10 2016).

58. Chang, C. C., Chow, C. C., *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4,** s13742–015–0047–8 (1 2015).

59. Shriner, D., Adeyemo, A., *et al.* Joint Ancestry and Association Testing in Admixed Individuals. *PLOS Computational Biology* **7,** e1002325 (12 2011).

60. Plummer, M., Best, N., *et al.* CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News* **6,** 7–11 (2006).

Supplementary material

Penn Medicine Biobank Team and Contributions

Leadership

Daniel J. Rader, M.D., Marylyn D. Ritchie, Ph.D., Michael D. Feldman M.D.

Contribution: Contributed to securing funding, study design and oversight.

Patient Recruitment and Regulatory Oversight

JoEllen Weaver, Nawar Naseer, Ph.D., M.P.H., Afiya Poindexter, Ashlei Brock, Khadijah Hu-Sain, Yi-An Ko

Contributions: JW manages patient recruitment and regulatory oversight of study. NN manages participant engagement, assists with regulatory oversight, and researcher access. AP, AB, KH, YK perform recruitment and enrollment of study participants.

Lab Operations

JoEllen Weaver, Meghan Livingstone, Fred Vadivieso, Ashley Kloter, Stephanie DerOhannessian, Teo Tran, Linda Morrel, Ned Haubein, Joseph Dunn

Contribution: JW, ML, FV, SD conduct oversight of lab operations. ML, FV, AK, SD, TT, LM perform sample processing. NH, JD are responsible for sample tracking and the laboratory information management system.

Clinical Informatics

Anurag Verma, Ph.D., Colleen Morse, P.T, D.P.T, M.S, Marjorie Risman, M.S., Renae Judy, B.S.

Contribution: All authors contributed to the development and validation of clinical phenotypes used to identify study subjects and (when applicable) controls.

Genome Informatics

Anurag Verma Ph.D., Shefali S. Verma, Ph.D., Yuki Bradford, M.S., Scott Dudek, M.S., Theodore Drivas, M.D., PH.D.

Contribution: A.V., S.S.V. are responsible for the analysis, design, and infrastructure needed to quality control genotype and exome data. Y.B. performs the analysis. T.D. and A.V. provides variant and gene annotations and their functional interpretation of variants.

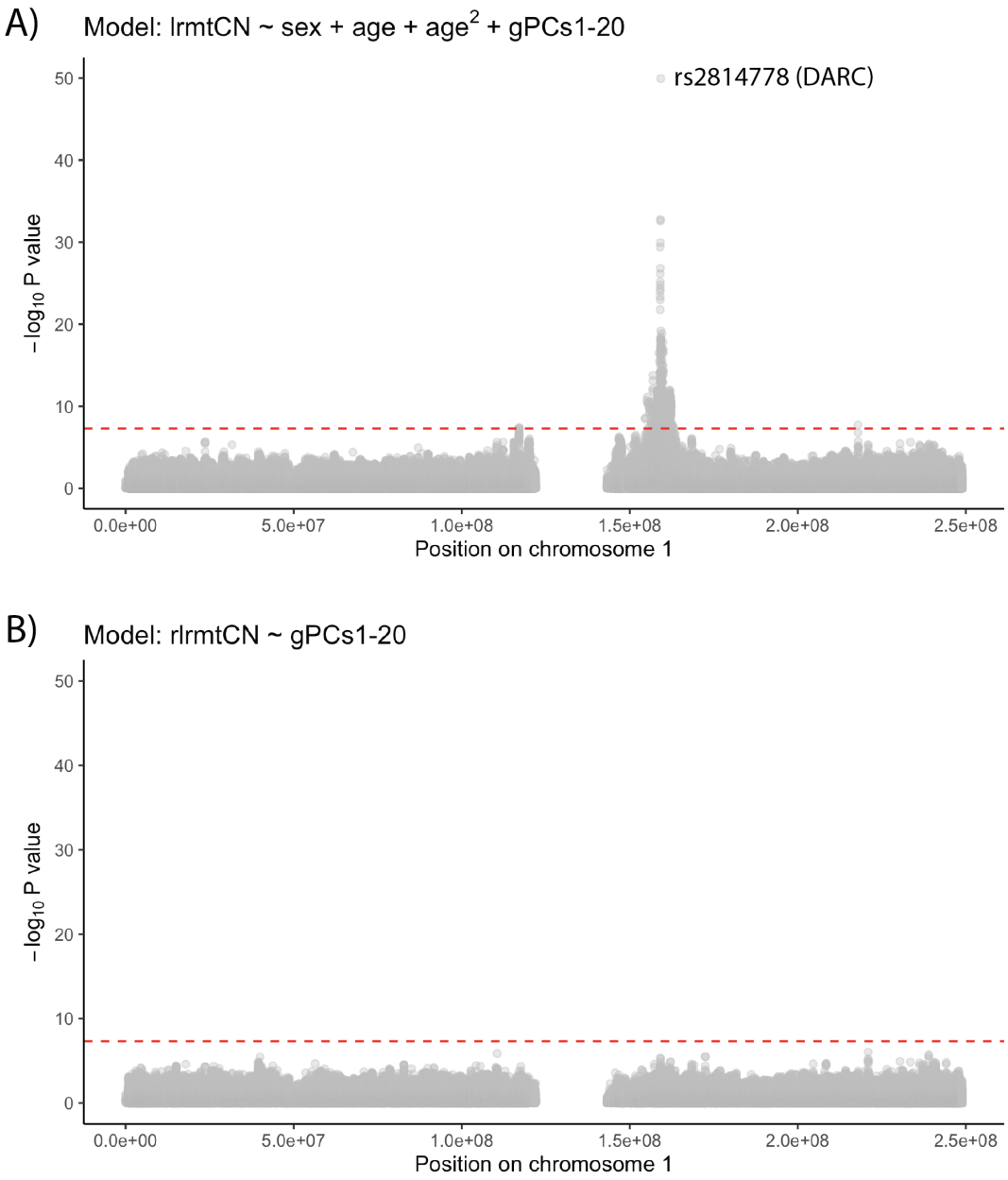

Figure S1: Manhattan plot of GWAS on rmtCN (A) before and (B) after correction for blood cell counts in the AFR cohort. Only chromosome 1 is displayed.

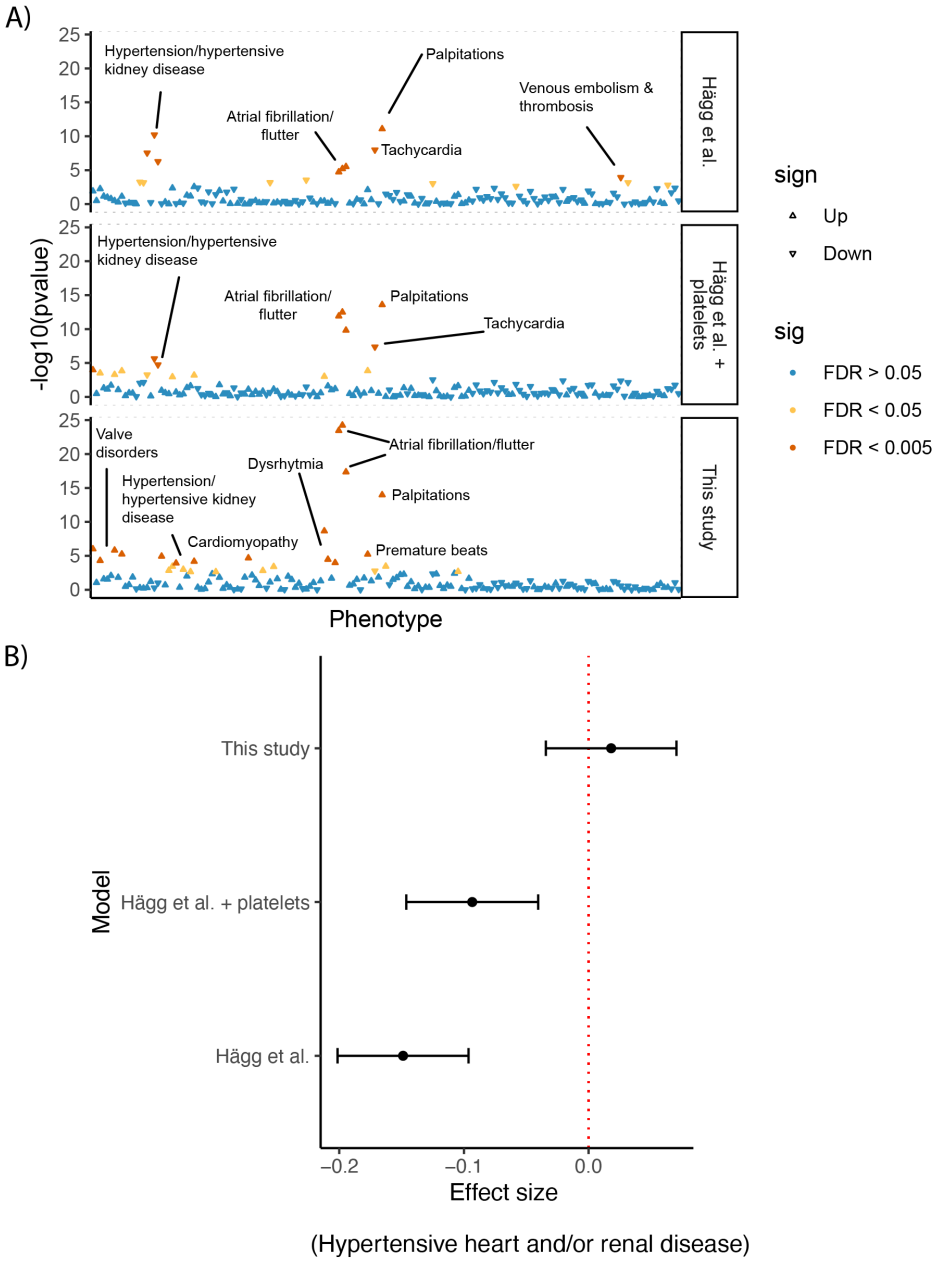

Figure S2: Sensitivity of association of mtDNA copy number with cardiac phenotypes to model choice. (A) The x-axis shows phenotypes arranged in order of phecode number such that similar phenotypes cluster together, and the y-axis shows the negative log of the association p-value. (Top panel) A model closely mimicking that used by Hägg *et al.* where, in addition to lrmtCN, we include sex, age, age^2, neutrophil %, lymphocyte %, total white blood cell count, and 20 PCs as predictors. (Middel panel) The same as previous model except for the addition of platelet count as a covariate. (Bottom panel) The model used in this study where we use rlrmtCN (residuals from the model described in the main text), sex, age, and age^2, and 20 PCs as predictors. (B) Forest plot illustrating the change in effect size for one phenotype (hypertension and renal disease) in the EUR cohort.

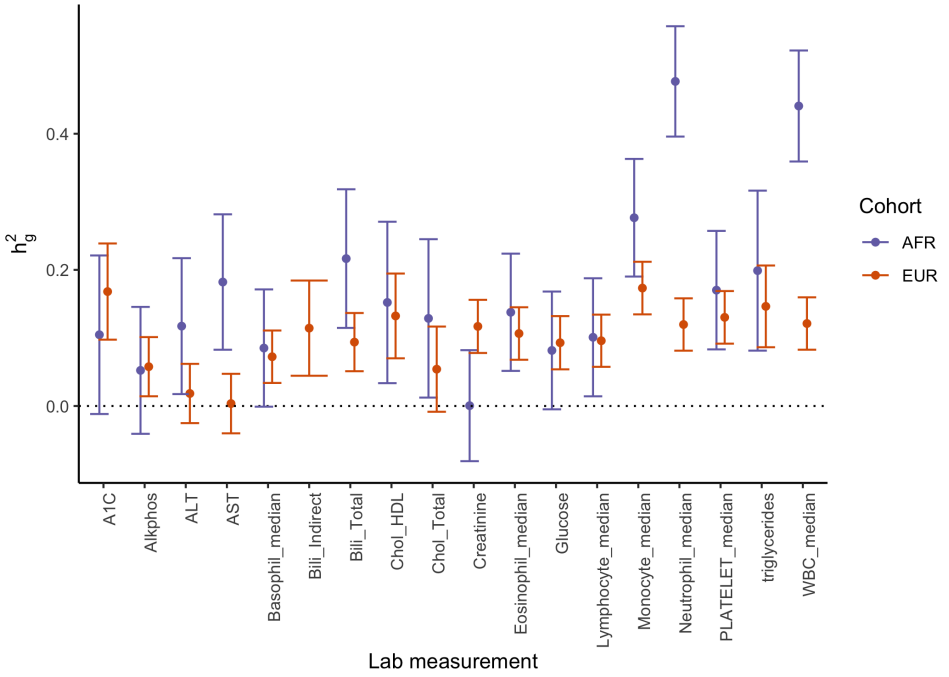

Figure S3: The heritability of lab measurements in the PMBB shown separately for the AFR and EUR cohort. Only lab measurements where the lower bound of the 95% CI was greater than 0 in at least in one of the cohorts is shown

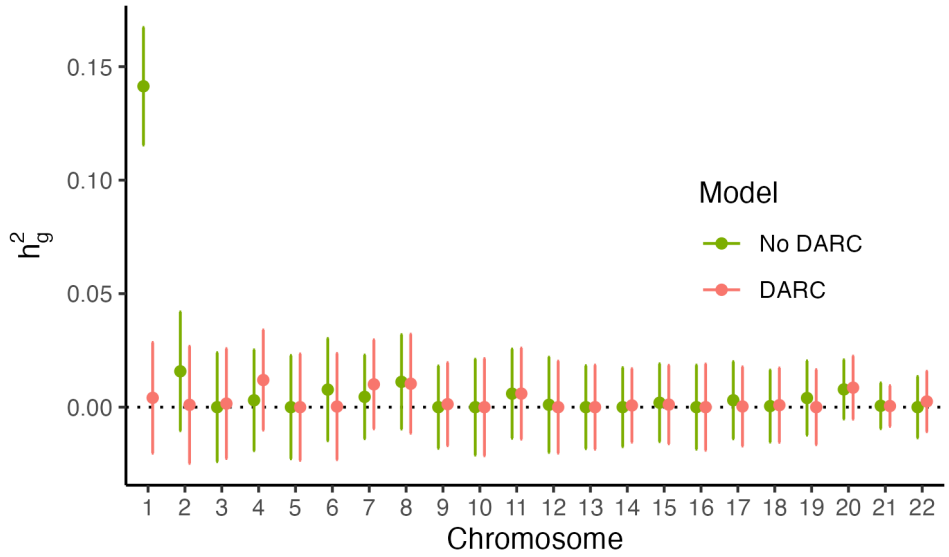

Figure S4: Heritability of neutrophil count partitioned by chromosome in the AFR cohort. The colors represent two models with and without genotype at the Duffy-null allele as a covariate. Both models included sex, age, age^2, and 20 PCs as covariates.

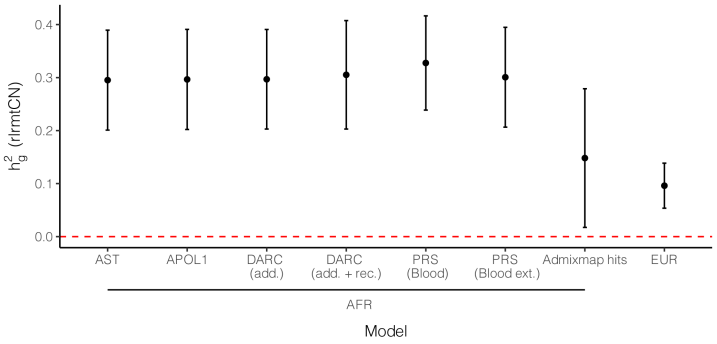

Figure S5: SNP heritability of rlrmtCN in the AFR (first 7 columns) and EUR (last columns) cohorts estimated using GCTA. All models included 20 genetic PCs calculated separately in each cohort. For the AFR cohort, the heritability was estimated with additional covariates: AST = amino aspartate transferase levels; APOL1 = genotype at the APOL1 locus; DARC (add.) = genotype at the rs2814778 SNP coded additively; DARC (add. + rec.) = additive and recessive coding for rs2814778; PRS (Blood) = polygenic risk scores for blood counts which were measured in the PMBB; PRS (Blood ext.) polygenic risk scores for an extended set of blood traits (see Methods for details).

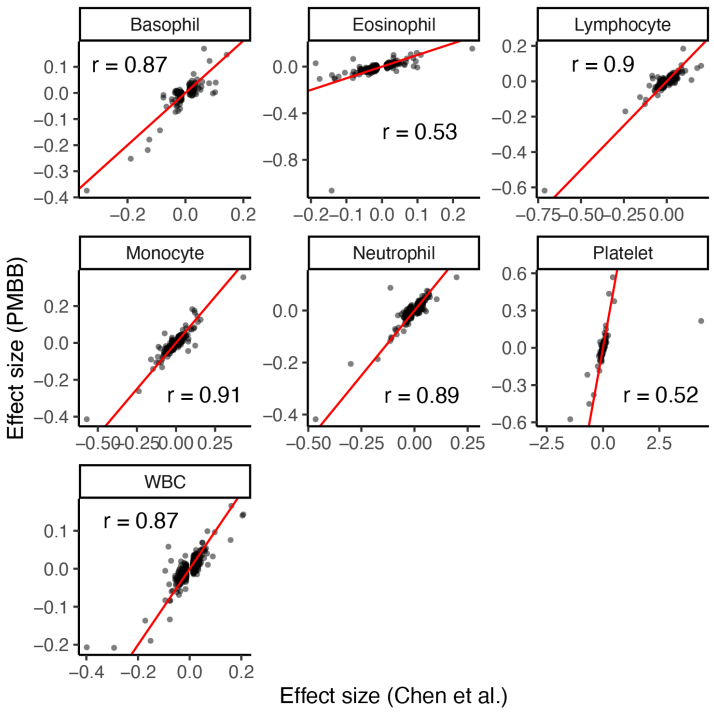

Figure S6: Effect sizes for variants discovered in Chen *et al.* [28] are correlated with their effects estimated in the PMBB EUR cohort. The red line represents $y = x$. The effects could only be re-estimated for traits which were available in the PMBB. One variant which can be seen as having a large effect size on platelet counts as estimated by Chen *et al.* was removed (see Methods for more details). The numbers in each plot show the correlation coefficients.

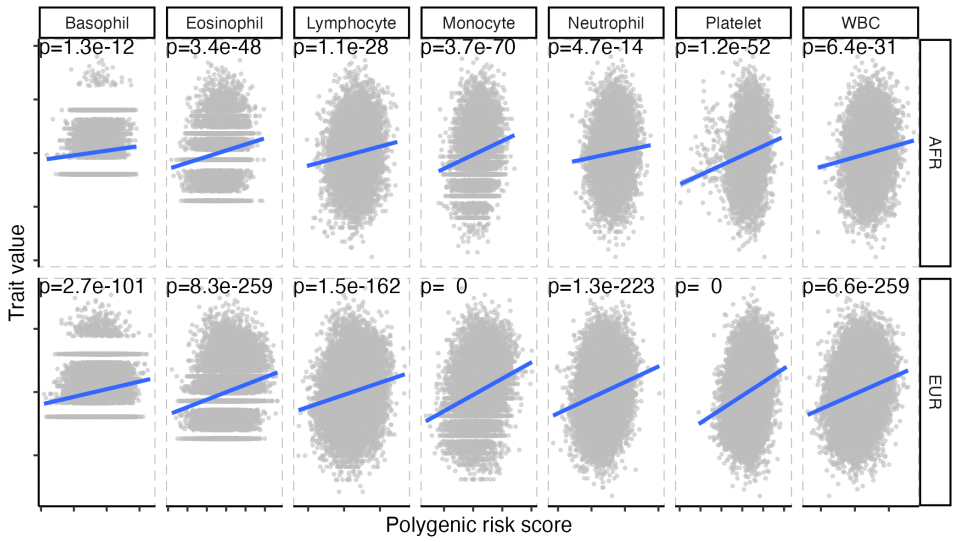

Figure S7: Polygenic risk scores (PRS) constructed using effects from Chen *et al.* [28] are strongly correlated with the actual phenotypes in both PMBB cohorts. The blue line represents the linear regression line.

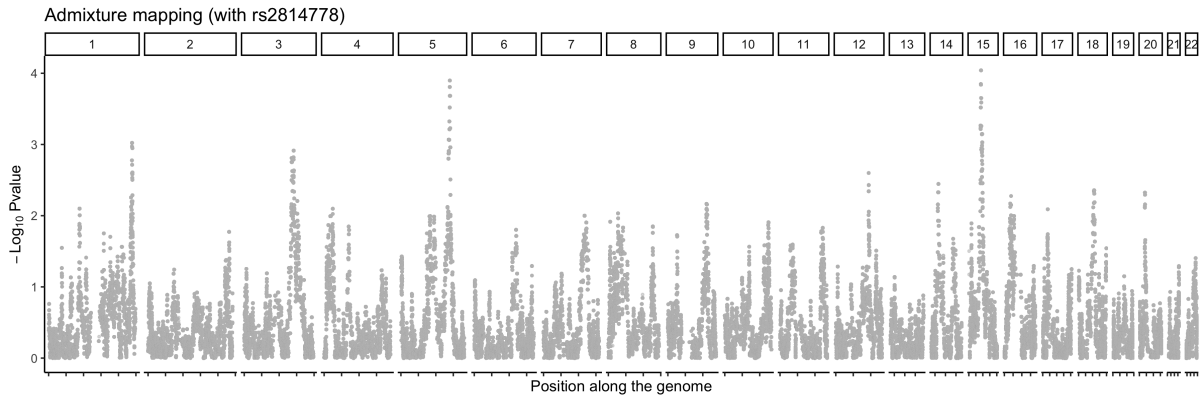

Figure S8: Admixture mapping of rlrmtCN in the AFR cohort. The x-axis shows the position along the genome and the y-axis shows the -$\log_{10}$ of the p-value of association between local ancestry at each position and rlrmtCN. Global ancestry proportion and the Duffy-null genotype were included as covariates.

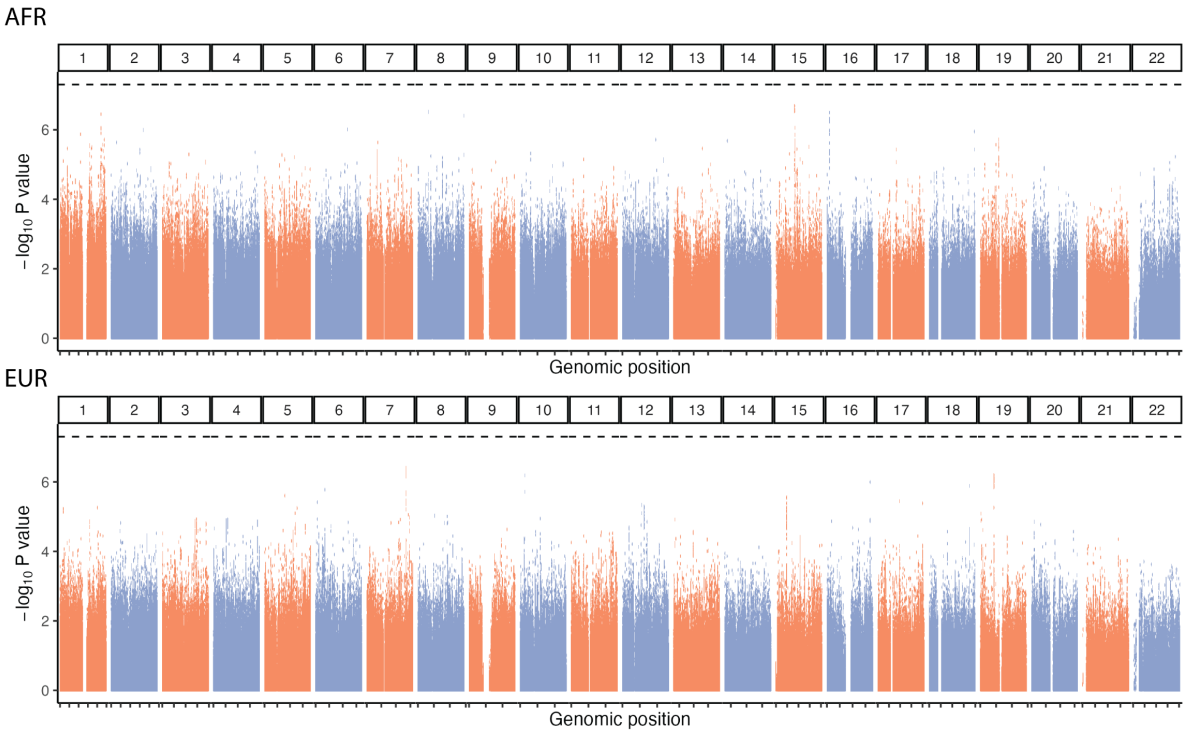

Figure S9: GWAS of mtDNA copy number (rlrmtCN) carried out separately in the AFR and EUR cohorts. The x-axis shows the genomic position, grouped by chromosomes (vertical panels) and the y-axis shows the -$\log_{10}$ of the association p-value. The dotted horizontal line represents the genome-wide significance threshold of 5×10^{-08}. The first 20 PCs, computed separately within each cohort, were included as covariates.

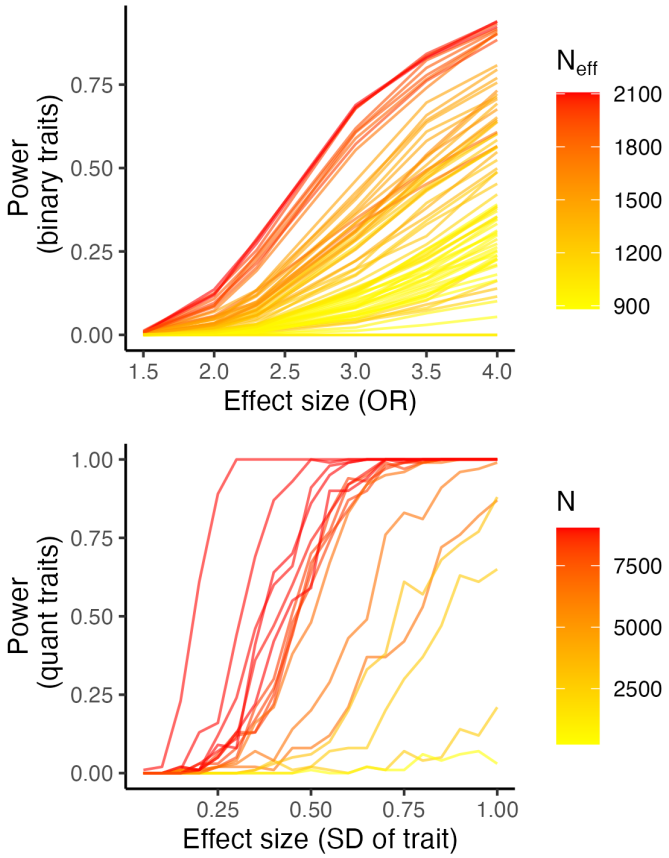

Figure S10: Power to detect a significant interaction effect between mitochondrial and nuclear ancestry for binary traits (case/control data) and quantitative traits (e.g. lab measurements). The x-axis lists the effect size, i.e., odds ratio (OR) or in units of standard deviation, for binary and quantitative traits, respectively and the y-axis shows the power of detecting an interaction effect at $\alpha = 5 \times 10^{-05}$. For quantitative traits, the color represents the sample size and for binary traits, it represents the effective sample size (N_{eff}): $n\phi(1-\phi)$ where ϕ is the proportion of cases and n is the sample size.

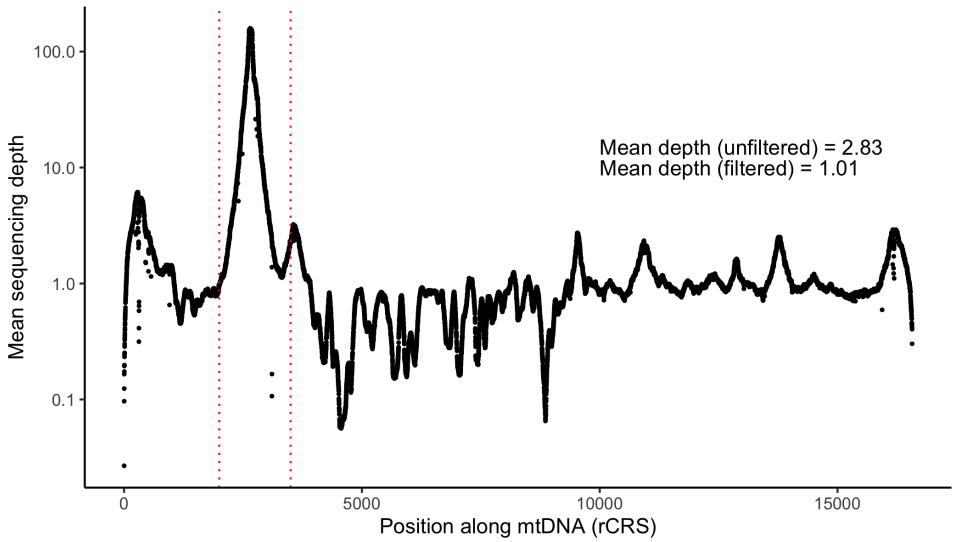

Figure S11: Mean sequencing depth (across individuals in the PMBB) of off-target reads aligning to the Revised Cambridge Reference Sequence (rCRS) of human mtDNA. Note that the y-axis is on a log-scale. Depth values from the region between the dotted red lines were filtered out for subsequent analysis.