Comprehensive Analysis of Mitochondrial DNA Variation in the Taiwan Biobank: Implications for Complex Traits and Population Genetics

Ting-Hsuan Chou[1], Pei-Miao Chien[1], Pin-Xuan Chen[1], Ni-Chung Lee[2,3,4], Hurng-Yi Wang[1,5,6,7], Yi-Cheng Chang[1,8,9], Pei-Lung Chen[1,2,5,6], Jacob Shu-Jui Hsu[1*]

[1]Graduate Institute of Medical Genomics and Proteomics, National Taiwan University College of Medicine, Taipei, Taiwan
[2]Department of Medical Genetics, National Taiwan University Hospital, Taipei, Taiwan
[3]Department of Pediatrics, National Taiwan University Hospital, Taipei, Taiwan
[4]Department of Pediatrics, National Taiwan University College of Medicine, Taipei, Taiwan
[5]Graduate Institute of Clinical Medicine, College of Medicine, National Taiwan University, Taipei, Taiwan
[6]Genome and Systems Biology Degree Program, National Taiwan University and Academia Sinica, Taipei, Taiwan
[7]Institute of Ecology and Evolutionary Biology, National Taiwan University, Taipei, Taiwan
[8]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan
[9]Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan
*Corresponding author

This study presents a comprehensive analysis of mitochondrial DNA (mtDNA) variation in the Taiwan Biobank (TWB), providing insights into the genetic characteristics of the Taiwanese population and the implications of mtDNA in complex traits. We performed mtDNA genotyping on 1,492 individuals using whole-genome sequencing data and imputed mtDNA variants for 101,473 participants from microarray data. Our analysis identified 23 confirmed pathogenic mtDNA variants, with approximately 1 in 180 individuals carrying such variants. Further exploration of mtDNA haplogroups and ancestry revealed no direct correlation between nuclear and mitochondrial genomes, which reflects their distinct inheritance patterns and evolutionary histories. In a mitochondrial genome-wide association study across 86 traits and 306 mtDNA variants, we discovered novel associations between *MT-ND2* gene variants and high myopia, as well as 14 mtDNA variants linked to renal function biomarkers. Notably, renal-associated variants clustered into two main groups: ancestral variants of macrohaplogroup M associated with poorer renal function and variants of the B4b sub-haplogroup linked to improved renal function markers. Our findings highlight the importance of population-specific genetic studies, contributing to our understanding of mitochondrial genetics in the Taiwanese population and its implications for health and disease.

Introduction

Mitochondria, often referred to as the "powerhouse" of cells, are double-membrane organelles presented in almost every eukaryotic cell and serve as the primary source of cellular energy production through ATP synthesis via oxidative phosphorylation (OXPHOS)[1]. Each mitochondrion contains its own genome (mtDNA), a circular double-stranded DNA molecule. In humans, the mitochondrial genome spans 16,569 base pairs and contains 37 genes encoding 13 protein subunits of the OXPHOS system, 22 tRNAs, and two rRNAs, all of which are crucial for its maintenance and expression[2,3].

Mitochondrial DNA exists in multiple copies within each cell, with the number of copies varying among individuals, tissues, and cells based on metabolic demands[4]. Due to multiple copies of mtDNA, accumulated variants can manifest in two forms: homoplasmy and heteroplasmy. Homoplasmy occurs when a specific variant is uniformly present across all mtDNA copies within an individual, while heteroplasmy arises when a mixture of different molecules exists. Since mtDNA is maternally inherited[5] and lacks intermolecular recombination[6], variants within a population are grouped into mtDNA haplotypes, commonly called haplogroups. These haplogroups have proven invaluable for tracking human biogeography[7].

Given the central role of mitochondria in energy production, it is not surprising that mtDNA variants have been associated with various diseases, including type 2 diabetes [8], cardiomyopathy[9], renal disease[10], and Parkinson's disease[11]. Recent population-level studies have further elucidated the influence of mtDNA variants on human health. For instance, research conducted within Biobank Japan involving 1,928 individuals revealed pleiotropy of mtDNA genetic risk on the five late-onset human complex traits such as creatine kinase[12]. Similarly, a larger investigation using the UK Biobank with 358,916 participants identified new associations between mtDNA variants and various traits, including type 2 diabetes and markers of liver and kidney function[13]. These studies underscore the critical role of common mtDNA variations in shaping complex human traits.

The Taiwan Biobank (TWB), established in 2012, is a pivotal genetic research resource. This prospective study has recruited over 200,000 individuals, generating a comprehensive repository of genomic and phenotypic data. The extensive data encompasses participants' demographics, socioeconomic status, family history, and self-reported disease profiles. Given that the majority of Taiwanese participants are of Han Chinese ancestry, analyses from TWB can offer unique insights into the mtDNA impacts on health in Taiwanese and other East Asian populations[14,15].

In this study, we comprehensively dissect the mitochondrial genetic profiles of the Taiwanese population utilizing data from the TWB. Our investigation had three primary objectives. First, to characterize the spectrum of mtDNA variation in the TWB, we implemented the GATK mitochondrial variant calling pipeline to accurately identify both homoplasmies and heteroplasmies from 1,492 WGS samples. Second, to explore the mitochondrial genetic structure and ancestry patterns, we performed Principal Component Analysis (PCA) on mtDNA variants and assessed the correlation between haplogroup classifications and their ancestral origins. Finally, to investigate the role of mitochondrial variation in complex traits, we accurately imputed mtDNA variants for 101,473 individuals using microarray data by building a Taiwanese-specific mitochondrial reference panel based on WGS data. This imputation enabled us to conduct a hypothesis-free mitochondrial-wide association study

involving 306 mtDNA variants and 81 complex traits. Through the analysis, we aim to uncover significant associations between mtDNA variations and various health outcomes, thereby contributing valuable insights into the role of mitochondrial genetics in complex traits within the Taiwanese population.

Methods

Data source

The research cohort was drawn from the Taiwan Biobank (TWB), a prospective study that has recruited over 200,000 individuals. The TWB provides extensive phenotype data, including demographics, socioeconomic status, environmental exposures, lifestyle factors, dietary habits, family history, and self-reported diseases gathered through structured questionnaires. Additionally, anthropometric measurements, including blood and urine samples, were obtained at the time of enrollment for subsequent biomarker analysis. For this study, we utilized a total of 1,492 whole-genome sequencing (WGS) data and 120,163 microarray data sourced from the TWB with ethical approval (TWBR11106-05 and 202108074RINC).

Mitochondrial DNA variant calling from the WGS data

In the TWB dataset, 1,492 WGS samples were sequenced by using HiSeq 2500 (n=555), HiSeq 4000 (n=634), or NovaSeq 6000 (n=303), which achieved high depth on autosomal chromosomes (35-45x). To identify mtDNA variants, we employed the GATK Best Practices for SNP/Indel Variant Calling in Mitochondria (V.4.1.8.0)[16]. In brief, the reads were aligned to GRCh38 using BWA-MEM (version 0.7.17), which includes the revised Cambridge Reference Sequence (rCRS, NC_012920.1) serving as the mitochondrial reference genome. A double alignment strategy was conducted to align the control region and the other region separately; this approach was necessary due to the circular nature of the mitochondria's genome. The caller used Mutect2 in mitochondria mode to detect variants with low variant allele frequency (VAF).

Subsequent stringent filtering steps were applied following GATK best practices for mitochondrial variants. At the sample level, we filtered samples with low mitochondrial copy numbers (<50) and excluded samples with contamination greater than 0.02. For genotype calls, we filtered out variants with a VAF of less than 0.1. At the variant level, we removed alleles found in regions where the sequence context makes it difficult to distinguish true variants from technical artifacts, including artifact-prone regions in six mtDNA positions (301, 302, 310, 316, 3107, 16182) and indels were only present as multi-allelic calls across all samples, as well as alleles for which no sample had a pass genotype.

Haplogroup assignment

We utilized high-quality mtDNA variants from WGS as input and classified each individual into a mitochondrial haplogroup by HaploGrep (v2.4.0)[17] based on the revised tree Phylotree17_FU1[18]. The first letter of the haplogroups was defined as macrohaplogroup. Each haplogroup is phylogenetically associated with three origins: African, Asian, and European, as described in MITOMAP[19].

Principal Component Analysis (PCA) on the nuclear DNA (nDNA) and mtDNA variant

To explore clustering patterns of autosomal and mitochondrial genetic structure, we employed Principal Component Analysis (PCA) as a linear dimensionality reduction technique. The autosomal variant detection was conducted as previously described[20]. Individuals with missing rates >0.02 were

excluded from the analysis. Nuclear PCAs were analyzed, focusing on a subset of SNVs adhering to the following criteria: SNVs with a VQSR tranche <99.7 and genotype call rate >0.98 biallelic variants. We then conducted PCA analysis using PLINK to elucidate the underlying clustering pattern[21].

Information regarding individuals' places of origin was obtained from a self-reported questionnaire, which included data on both maternal and paternal ancestries. Based on this data, individuals from the TWB were classified into four main clusters: "Holo," "Hakka," "Southern Han Chinese," and "Northern Han Chinese." The categorization of "Southern Han Chinese" and "Northern Han Chinese" was based on their respective provincial regions in relation to the Yangtze River.

Quality control of mtDNA variants from the microarray data

We utilized 120,163 microarray data from the TWBv2 SNP array, where participants were genotyped using the Thermo Fisher Scientific Axiom Genome-Wide TWB 2.0 Array. This genotyping array includes 752,921 probes designed to assay 686,463 SNVs, as well as 815 mtDNA variants. The analysis was performed using officially released PLINK files, which were processed with the Axiom Analysis Suite, following the manufacturer's recommended best practice workflow.

To ensure high-quality mtDNA variant data, we applied stringent quality control (QC) criteria. We excluded individuals with a sample missing rate >0.02 and variants with a missing rate >0.02. Additionally, we compared allele frequencies (AFs) observed in the microarray with those from WGS. Variants with AFs deviating more than 0.2 from WGS observed frequencies were removed from the subsequent analysis.

To evaluate the detection limit of heteroplasmies under standard procedures, we leveraged data from 1,426 individuals who had both WGS and microarray data available. We compared the genotype calls between the array and WGS datasets for variants that had designed probes and were detected in the WGS datasets to assess consistency. Specifically, we used the WGS call set as the truth set, allowing us to compare the detection rates of the array data across different VAF bins with those detected by WGS. This analysis provides a thorough evaluation of the microarray's capability to detect heteroplasmies under different VAFs. As depicted in Figure S4B, since the detection of heteroplasmies is rare in the genotyping of mtDNA variants, we set all the heteroplasmies as missing in WGS when constructing the mitochondrial imputation reference panel.

Evaluation of imputation accuracy

Before initiating imputation, we assessed the feasibility and accuracy of imputing mtDNA variants using a sample of 1,426 individuals who had both WGS and microarray data available. These individuals were divided into two groups: 1,000 formed the reference set, and the remaining 426 constituted the test set. We then constructed an mtDNA imputation reference panel from the WGS sequences of the 1,000 individuals and imputed mtDNA variants from the microarray data of the 426 test individuals.

The mtDNA scaffold haplotypes were pre-phased using SHAPEIT2[22] before imputation to ensure compatibility with the IMPUTE2 algorithm. Imputation of mtDNA variants was conducted by IMPUTE2[23], covering the full mitochondrial genome with region boundaries set to -int 1 16579. Given

the absence of an available mitochondrial genetic map, we created a genetic map indicating little to no recombination of the mitochondrial genome.

We evaluated the imputation performance in both haploid and diploid settings and observed no significant differences. After imputation, we assessed genotype concordance between the WGS and imputed variants for each AF bin separately (0-0.005, 0.005-0.01, 0.01-0.05, 0.05-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4, 0.4-0.5, 0.5-0.6, 0.6-0.7, 0.7-0.8, 0.9-1.0). This analysis enabled us to establish post-imputation filtering criteria, setting a threshold INFO score of 0.7 to ensure high confidence in the imputed data.

Imputation of microarray data

We constructed a Taiwanese-specific mitochondrial reference panel from 1,465 WGS sequences. After applying pre-imputation filtering, our dataset included 101,473 unrelated EAS samples and 563 mtDNA variants that were biallelic, had a missing rate of less than 0.02, and exhibited AF deviations from WGS of less than 0.2. Post-imputation QC filtering was applied to variants with MAF >0.01 and an imputation INFO score >0.7. This process resulted in 306 high-quality mtDNA variants, which demonstrated high allele frequency consistency with WGS data (Pearson r = 0.999), suitable for subsequent association analysis.

Preparation of phenotypic traits for analyses

We gathered phenotypic data on disease status via questionnaires and clinical measurements from the TWB. We included self-reported diseases with at least 100 cases, resulting in a selection of 48 diseases. For quantitative traits, these measurements encompass seven categories, offering a comprehensive assessment of participant health. These categories include anthropometric measurements (n = 7), lung function (N=3), bone density (n=2), cardiovascular function (n=3), hematological parameters (n=5), metabolism (n=6), liver (n=6) and kidney function (n=6). We removed individuals whose measurements deviated by more than five standard deviations from the mean to eliminate outliers and applied rank inverse normal transformation (RINT) to standardize each phenotype.

Mitochondrial genome-wide association analysis

Initially, we defined a subset of unrelated East Asian (EAS) individuals to be included in the downstream association analysis. We inferred a genetically EAS group from the TWB using 1000 Genomes (1KG) phase 3 samples as the population reference panel[24]. PCA was performed on common autosomal variants between TWB and 1KG datasets. The variants selected were biallelic with a missing rate <0.02, r^2 <0.1, MAF >0.05, and located outside long-range LD regions (chr6: 25-35Mb, chr8: 7-13Mb). Then, we used the Random Forest classifier to predict the EAS group[15].

To ensure the analysis was conducted on unrelated individuals, we calculated kinship coefficients using KING[25] based on autosomal variants. Individuals identified with a kinship coefficient greater than 0.0884, indicative of second-degree relatives or closer relationships, were excluded from the study.

For the statistical analysis, we implemented an additive model and conducted regression analyses using the glm() function in R, tailored for quantitative traits (linear model) and binary traits (logistic model).

These analyses were adjusted for age, sex, age and sex interaction, the top 10 nuclear PCs, and genotyping batches to mitigate potential confounding factors. For sex-specific phenotypes, adjustments were made for age and the top 10 nuclear PCs. We applied a Bonferroni correction to set the significance threshold at $P = 0.05/306 = 1.4 \times 10^{-4}$. Additionally, considering the number of traits analyzed, we adopted more conservative thresholds of $P = 0.05/306/48\ traits = 3.4 \times 10^{-6}$ for binary traits, and $P = 0.05/306/38\ traits = 4.3 \times 10^{-6}$ for quantitative traits.

Association of mtDNA haplogroups

Using imputed mtDNA genotypes, we applied Haplogrep2 to assign haplogroup for each individual [17]. We applied filters before performing haplogroup assignments, including individual missing rate <0.02, genotype missing rate <0.02, and MAF >0.01. Following the haplogroup assignment, we analyzed the distribution of haplogroups and conducted PCA analysis on these mtDNA variants. The results from the genotyping array were compared with those in WGS to assess the similarity and accuracy.

Subsequently, we focused on the association between haplogroups and renal markers, including serum creatinine (Scr) and estimated glomerular filtration rate (eGFR). Both values were rank-inverse normal transformed. The association analysis was restricted to haplogroups represented by at least 500 individuals to ensure sufficient statistical power. We employed GLM with a Gaussian distribution, adjusting for genotyping batches, age, and sex, and the most prevalent haplogroup M7b was set as the reference. We set the significance threshold at $P < 1.22 \times 10^{-3}$ $(0.05/41\ haplogroups)$, calculated based on the number of haplogroups tested.

Results

mtDNA genotyping across 1,492 TWB individuals

In this study, we conducted mtDNA genotyping on 1,492 individuals from the TWB using the GATK pipeline for variant calling. This approach allowed us to characterize the spectrum of mtDNA variation in the TWB by identifying both homoplasmic and heteroplasmic derived from WGS data. Following rigorous filtering criteria, we retained 1,465 samples for further analysis. These samples exhibited nDNA coverage ranging from 35x to 45x, and mtDNA coverage ranged from 2500x to 3000x. The average mtDNA copy number across these samples was estimated to be 155[26]. (Figure 1A-1C).

Our analysis unveiled a total of 2,361 mitochondrial variants, consisting of 2,289 SNVs and 72 indels. The majority of SNVs were transitions rather than transversions, and most detected variants were found to be homoplasmic. Furthermore, our investigation of AF spectra revealed that over half of the identified variants occurred uniquely in one or two individuals, with the majority having AFs below 1% (Figure S1).

When comparing our TWB mtDNA variant call set with the gnomAD database (v3.1), the Taiwanese cohort exhibited a close genetic resemblance to gnomAD's East Asian population, with a Pearson correlation of 0.998 for AF distributions of shared variants (Figure 1D and IE). However, about 30% of the variants found in TWB were not reported in the East Asian population of gnomAD, suggesting notable differences and underscoring the importance of population-specific studies.

Capability of mtDNA and nuclear genetic structure in reflecting ancestral information

Due to the nature of maternal inheritance and lack of recombination, mtDNA haplogroups provide a framework for understanding the maternal lineage[7]. The Phylotree database offers a thorough phylogenetic tree of worldwide human mtDNA variation, comprising over 5,400 haplogroups with their defining mutations[27]. To discern the haplogroup within each individual, we used HaploGrep2 for haplogroup assignment[17]. Our analysis revealed that the most prevalent mtDNA haplogroups among the TWB participants were M, D, F, and B, all indicative of Asian ancestry and reflective of the genetic background of the Taiwanese population (Figure 2A and 2B). In the examination of variants presented in each mitochondrial haplogroup, we found that mutation count correlates with the haplogroups' evolutionary lineage, which was also supported by earlier findings[16,28] (Figure S2).

In contrast to the well-documented nuclear structure that reflects population structure within the Taiwanese population[14,15], less is known about mitochondrial genome structure in Taiwan. Moreover, to appropriately control for population stratification—essential for robust association studies—it is necessary to elucidate the correlation between two genetic structures. We conducted principal component analyses (PCA) on both nuclear and mitochondrial genomes. Nuclear PCA revealed a high degree of homogeneity among the TWB population, with individuals reporting Northern Han Chinese ancestry positioned peripherally in the genetic landscape (Figure 2C). In contrast, the mitochondrial PCA revealed a clear clustering of individuals based on their haplogroup assignments (Figure 2F), reflecting anticipated diversity in mitochondrial genotypes. However, a limited correlation was observed between nuclear PCs and mitochondrial haplogroups, and no substantial correlation was

evident between mitochondrial PCs and maternal ancestries (Figure 2D and 2E; Figure S3). These findings support the arguments made in the BBJ that there is no correlation between two genomes. Additionally, nuclear genome-derived PCs adequately account for the influence of population stratification when evaluating associations with mtDNA variants.

Prevalence of confirmed pathogenic mtDNA variants in TWB

To comprehensively the mitochondrial genetic profile, we investigated the prevalence of 96 confirmed pathogenic mtDNA variants listed in MITOMAP across 1,465 WGS samples[19]. We identified 6 pathogenic variants with a VAF greater than 10%, indicating a carrier frequency of 0.556%, which aligns with earlier estimates of about 0.5%[16,28]. To extend our findings and enhance the scope of our study, we included 120,163 TWBv2 microarray samples. The TWBv2 SNP array was specifically designed to cover a wide range of disease-relevant variants and serves as a valued source for this purpose. This genotyping array includes 815 mtDNA variants, with 58 confirmed pathogenic variants.

From the microarray, we identified 20 pathogenic variants, revealing a carrier frequency of about 0.487%. In total, we pinpointed 23 confirmed pathogenic mtDNA variants within the TWB (Table S2). Many of these variants are associated with mitochondrial diseases known for their incomplete penetrance, including aminoglycoside-induced hearing loss, Leber's Hereditary Optic Neuropathy (LHON) and mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes (MELAS)[29]. Among the identified variants, m.1555A>G (rs267606617), linked to aminoglycoside-induced hearing loss, exhibited the highest frequencies in both WGS and microarray data. Other variants include the identification of LHON-associated variants m.4171A>G (rs28616230), m.11778G>A (rs199476112), and m.14484T>C and a MELAS-associated variant m.3697G>A (rs199476122). These findings contribute to the understanding of mitochondrial pathology within the Taiwanese population.

Imputation of mtDNA variants

To explore the complex interplay between mtDNA variants and a range of complex traits, our study leveraged microarray data from the TWB. We imputed mtDNA variants from the microarray using a reference panel derived from the 1,465 WGS, enabling us to analyze mtDNA variants across 101,473 participants.

First, we rigorously validated the accuracy of genotyped mitochondrial variants in microarray by comparing AFs between shared mtDNA variants from the microarray and WGS data. We found strong concordance, except for one variant that showed a significant deviation, which was removed in the following analysis (Figure S4A). Next, we assessed the capability of the microarray data to detect low-level heteroplasmies. Utilizing a subset of 1,426 individuals with both WGS and microarray data, we evaluated whether genotype calls presented in WGS could be detected in microarray samples. Our results showed limitations of microarray in identifying heteroplasmic variants, particularly at heteroplasmy levels lower than 70% (Figure S4B). To address this, we excluded all heteroplasmies from the reference panel from the reference panel construction. Finally, to define the appropriate threshold for post-imputation quality control, we utilized the same subsets of 1,426 individuals and calculated the genotype concordance rate of imputed mtDNA variants within each VAF bin (Methods). Based on this analysis, we selected an INFO score >0.7 (estimated by IMPUTE2) to ensure high-

quality imputed genotypes (Figure S4C). With the imputation quality ascertained, we refined our dataset through a workflow tailored for mtDNA variant imputation (Figure S5). After applying this quality filter and restricting AFs to within 0.1 deviation from WGS data, the final dataset included 563 high-confidence imputed mtDNA variants.

Mitochondrial genome-wide association study

To comprehensively analyze the association between mtDNA variants and complex traits, we conducted mitochondrial genome-wide association analyses. We retained 306 mtDNA variants with an MAF greater than 0.01 for downstream analysis (Table S2). These variants demonstrated high allele frequency consistency with WGS data (Pearson r = 0.999).

Focusing on diseases prevalent among TWB participants, we included 48 binary traits for analysis (Table S3). Additionally, we leveraged 38 quantitative traits derived from blood biomarkers, excluding outliers and applying rank inverse-normal transformations as appropriate (Table S4; Methods). Using the imputed dosages (0 or 2) of the individuals, we conducted regression analyses with adjustment for sex, age, age and sex interaction, genotype batches, and potential population stratification by including the top 10 nucPCs as covariates (Figure S6).

Despite no mtDNA variants meeting the conservative significance threshold, several variants were observed to associate with traits at a lenient threshold adjusting for total mtDNA variant counts ($P = 0.05/306 = 1.4 \times 10^{-4}$) (Figure 3; Table S5 and S6).

Association of mtDNA variants with high myopia

In our study, we identified associations between two mtDNA variants and high myopia. The variants rs28570593 (m.5054 A>G) and rs367778601 (m.5147A>G) exhibited odds ratios (ORs) of 2.47 ($P = 1.63 \times 10^{-5}$) and 1.96 ($P = 1.08 \times 10^{-4}$), respectively, suggesting their potential links to the condition. Both variants are synonymous variants located on the *MT-ND2* gene, with the mere linkage between them (r^2=0.148) indicating they may independently contribute to high myopia (Figure 4). The first variant, rs28570593, showed a higher prevalence in cases (0.056) compared to controls (0.010). The second variant, rs367778601, also displayed a significant case-control frequency difference, with allele frequencies of 0.080 in cases versus 0.023 in controls. These findings are consistent with previous research that reported associations of different variants on the MT-ND2 gene with high myopia in the Han Chinese population.[30].

Association of mtDNA variants with renal function biomarkers

We identified 14 mtDNA variants associated with biomarkers indicating renal function. specifically serum creatinine (Scr) levels and estimated glomerular filtration rate (eGFR) (Figure 5A; Table 1). Due to the lack of recombination in the mitochondrial genome[31], we hypothesized that these variants may share common haplotypes. To investigate this, we first calculated pairwise linkage disequilibrium (LD) r^2 focused on common mtDNA variants (MAF >0.05) using PLINK[21]. The results indicated multiple common haplotypes spanning the entire genome (Figure S7).

To further elucidate whether mtDNA variants associated with renal function belong to the same haplotypes, we categorized these variants based on their LD r^2, revealing two distinct clusters (Figure 5B). The first set primarily comprises ancestral markers for the divergent mitochondrial super-haplogroups M and N. Variants defining super-haplogroup M are associated with increased risk factors for impaired renal function, including higher Scr levels and lower eGFR. Conversely, alleles commonly found in super-haplogroup N correlate with decreased Scr levels and higher eGFR, suggesting a potentially protective effect on kidney function. The second set of variants belongs to haplogroup B4b, a subgroup within the N lineage. These B4b variants demonstrate associations with lower Scr levels and higher eGFR, further supporting a protective effect against renal impairment.

B4b haplogroup is associated with renal function

Our results suggest that a specific haplogroup background can affect susceptibility to renal function. To investigate this further, we conducted an analysis of haplogroup-specific effects on renal function markers. We utilized imputed and quality-controlled mtDNA variants to assign haplogroups to individuals using HaploGrep2[17], achieving a mean quality score of 0.906 (Figure S8). We then tested the association between sub-haplogroups and renal function markers, specifically Scr and eGFR. Sub-haplogroups were included in the analysis if they were carried by more than 500 individuals.

Our analysis revealed that the B4b sub-haplogroup was significantly associated with decreased Scr ($P = 8.15 \times 10^{-5}, \beta\ (\mathrm{SE}) = -0.06\ (0.01)$) and increased eGFR ($P = 7.33 \times 10^{-5}, \beta\ (\mathrm{SE}) = 0.05\ (0.01)$) (Figure S9; Table S7 and S8). These findings suggest a potentially protective role of the B4b haplogroup in renal function. However, when we conducted an association analysis between the B4b haplogroup and self-reported renal failure status, no significant association was observed. This discrepancy suggests that while B4b may have a beneficial effect on renal function markers, its influence may be subtle and not directly translate to a reduced incidence of clinically diagnosed renal failure.

Discussion

In this study, we comprehensively analyzed the mtDNA of individuals from the TWB. Our findings provide valuable insights into the genetic characteristics of the Taiwanese population and contribute to our understanding of mitochondrial genetics and its implications for health and disease.

Our mtDNA haplogroup analysis revealed a predominance of Asian-associated haplogroups (M, D, F, and B), aligning with the expected genetic heritage of the Taiwanese population. However, the distinct patterns revealed by mtDNA-based PCA compared to autosomal PCA suggest limitations in using mtDNA alone to capture ancestral variations within the TWB. Our findings indicate no direct correlation between nuclear and mitochondrial genomes, supporting similar observations from the BBJ study[12]. This contrasts with findings from the UKB, where associations between mitochondrial sub-haplogroups and nuclear PCs were identified[13]. This discrepancy highlights the complexity of genetic correlation and underscores the need for a more integrative approach to elucidate the full spectrum of associations between the two genomes in both genotype- and population-level in diverse populations.

We conducted a large-scale mitochondrial-genome-wide association study across 86 traits and 306 mtDNA variants in 101,473 Taiwanese participants. Our study demonstrated a successful imputation using WGS as a reference panel and revealed abundant mtDNA variants risk on complex traits. We identified associations between mtDNA variants and high myopia (HM). Specifically, two variants in the *MT-ND2* gene—rs28570593 and rs367778601—were linked to HM, particularly in individuals requiring spherical corrections of -10 diopters (D) or more. The *MT-ND2* gene encodes NAHD dehydrogenase 2, a component of complex I in the mitochondrial respiratory chain system, crucial for cellular energy production[32]. Previous studies have indicated *MT-ND2* variants, such as m.5244G>A[33] and m.4640C>A[34], in Leber hereditary optic neuropathy (LHON), suggesting a role of *MT-ND2* dysfunction in ocular diseases. In a recent study, Xing et al. identified nine novel mitochondrial variants associated with HM, including rs370378529 in *MT-ND2,* with an odds ratio (OR) of 5.25[30]. These findings underscore the pivotal role of complex I in cellular energy metabolism, where subtle dysfunctions may not cause overt disease but contribute to conditions like myopia.

mtDNA variants have emerged as significant factors in kidney function[35], reflecting the high mitochondrial content and oxygen demand of renal tissues[35]. In this study, we identified 14 mtDNA variants associated with renal function biomarkers, highlighting the potential relationship between mtDNA and kidney health. These mtDNA variants, while seemingly unrelated and distributed across the whole mitochondrial genome, can be categorized into two distinct sets: ancestral variants linked to the macrohaplogroup M diverges from L3, as well as the B4b sub-haplogroup. Notably, ancestral variants for macrohaplogroup M were associated with worse renal function indicators, with the missense variant m.10398A>G in the *MT-ND3* gene showing the most significant association with decreased eGFR. The G allele of this variant and the association of risk trends in renal function are reported in previous studies[36]. This variant is associated with impaired mitochondrial function, specifically affecting the function of Complex I of the electron transport chain, which is critical for energy-intensive processes in renal cells[32].

Conversely, the B4b sub-haplogroup was significantly associated with improved renal function markers, including decreased Scr and increased eGFR. While the haplogroup is predominantly found

in East and Southeast Asia[37], the finding underscores the value of conducting association studies in non-European cohorts to uncover population-specific genetic influences on renal function.

Comparing our findings with other large-scale studies, such as the UK Biobank and Japan Biobank[12,13,36], reveals both consistencies and population-specific differences in mtDNA variants' implications in renal function indicators. Most variants associated with serum renal indicators in these studies were either absent or not replicated in our TWB cohort, likely reflecting population-specific genetic influences on renal function. However, the variant m.3010G>A, located on the *MT-RNR2* gene, showed a consistent association with decreased renal function across studies, being linked to increased cystatin C, decreased eGFR^{cy}, and decreased eGFR^{crcy} in the UK Biobank and decreased eGFR in our cohort. This consistency across diverse populations strengthens the evidence for this variant's role in kidney function.

A recent study from the INTERVAL dataset demonstrated that common mtDNA variants influence blood *N*-formylmethionine (fMet) levels, a critical amino acid in mitochondrial translation. Variants in mtDNA haplogroups Uk and H4 were linked to elevated fMet levels, potentially disrupting mitochondrial protein synthesis and degradation. This disruption could impair kidney function and contribute to age-related diseases by altering energy metabolism and protein dynamics[38]**.** Notably, the m.10398A>G variant in our cohort, associated with decreased eGFR, also presented in the INTERVAL dataset and linked to increased fMet levels, providing a potential mechanistic explanation for its effect on renal function.

Despite its contribution, our study still has limitations. First, while we comprehensively dissected the mitochondrial genetic profiles in the Taiwanese population, our analysis predominantly identified SNVs and indels, but larger structural variants (SVs), which could be equally significant, remain undetected. Second, due to the constrained sample size in the association analyses, our focus was primarily on microarray data, thus limiting to homoplasmic or near-homoplasmic variants. This may overlook the role of heteroplasmic variants, which are crucial in certain diseases[39]. Moreover, our analysis relies on self-reported data from questionnaire surveys, which may introduce inaccuracies due to biases or errors in self-reporting. Finally, given that some mtDNA variants were not imputed due to the limited genotyped SNPs and the nature of common haplotypes spanning the whole mitochondrial genome, it is challenging to map the causal variants. While the study provides insights into the association between mtDNA variants and complex traits, further replication datasets or molecular studies are necessary to elucidate the mechanisms influencing disease processes.

Our analysis revealed that neither the mtDNA variants nor haplogroups we identified were associated with self-reported renal failure status. This observation suggests that the effects of these variants may be subtle and act as phenotype-modifiers[40], with their cumulative impact potentially increasing the risk of poor kidney outcomes over the long term. Thus, understanding the role of mitochondrial genetic background is crucial for gaining insights into phenotypic variability.

In conclusion, our study provides a comprehensive characterization of mtDNA within the TWB, significantly enhancing our understanding of mtDNA diversity and its implications for health and disease in the Taiwanese population. These findings emphasize the critical importance of including

diverse genetic backgrounds in mitochondrial research by revealing insights into genetic diversity and the pivotal roles of mtDNA variants in complex traits.

Figures

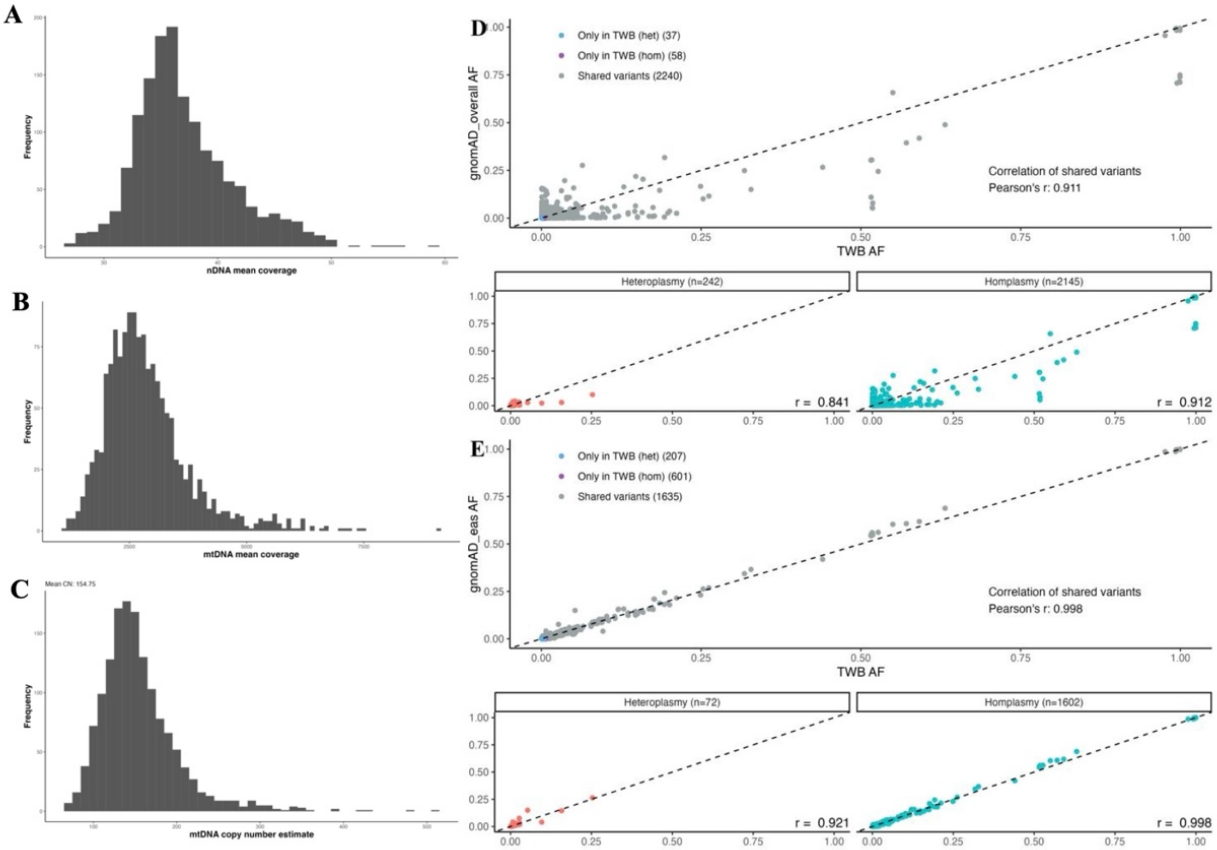


Figure 1. mtDNA variant statistics in Taiwan Biobank.
(A-C) The histograms display nDNA coverage (A), mtDNA coverage (B), and mitochondrial copy number (mtCN) (C) across 1,465 WGS samples.
(D) Scatter plots displaying correlations of allele frequencies between TWB and gnomAD: overall AFs (D), East Asian (EAS) AFs (E). Grey dots represent shared variants between the two datasets. Blue and purple dots denote heteroplasmies and homoplasmies exclusive to TWB. The line of identity is depicted in each plot, emphasizing the correlations. The subplots below compare heteroplasmic and homoplasmic variants separately.

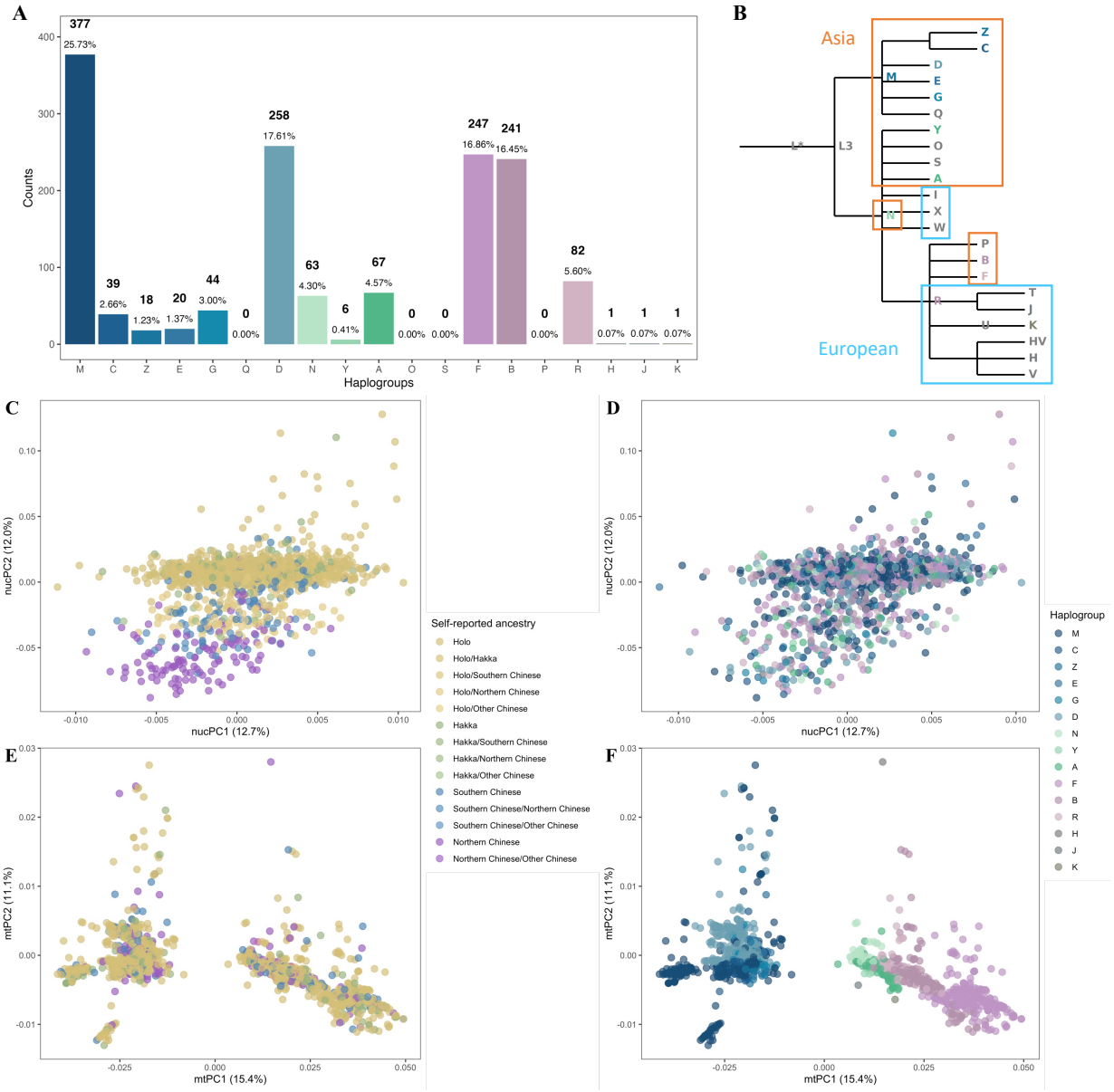


Figure 2. mtDNA haplogroups analysis in TWB

(A) The bar graph displays the prevalence of mtDNA haplogroups among TWB participants, arranged according to their relationships in the mitochondrial Phylotree. Haplogroups deriving from macrohaplogroup M are colored in shades of blue, those from N in shades of green, and those from R in shades of purple. For clarity, each haplogroup's color is consistent across all subfigures.

(B) A simplified tree depicting mtDNA haplogroups found among TWB participants, with each haplogroup uniquely colored. Asian and European haplogroups are delineated by distinct colors as defined in MITOMAP. Haplogroups not present in the TWB are shown in grey.

(C and D) PCA analyses using nuclear variants from 1,492 participants. Participants are colored by self-reported parental ancestry in (C) and haplogroup assignment in (D).

(E and F) PCA analyses using mtDNA variants from 1,492 participants. Participants are colored by self-reported maternal ancestry in (E) and haplogroup assignment in (F).

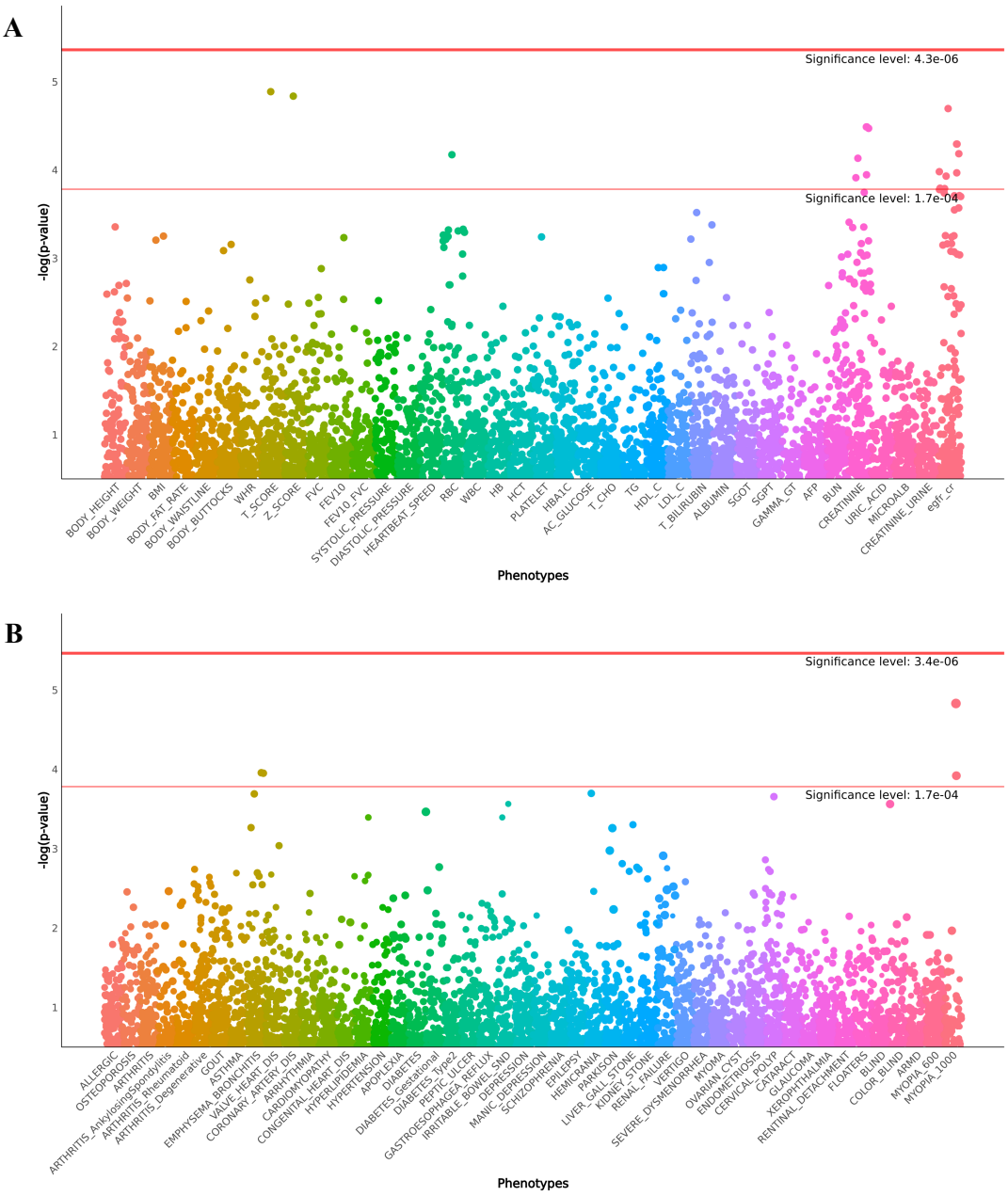
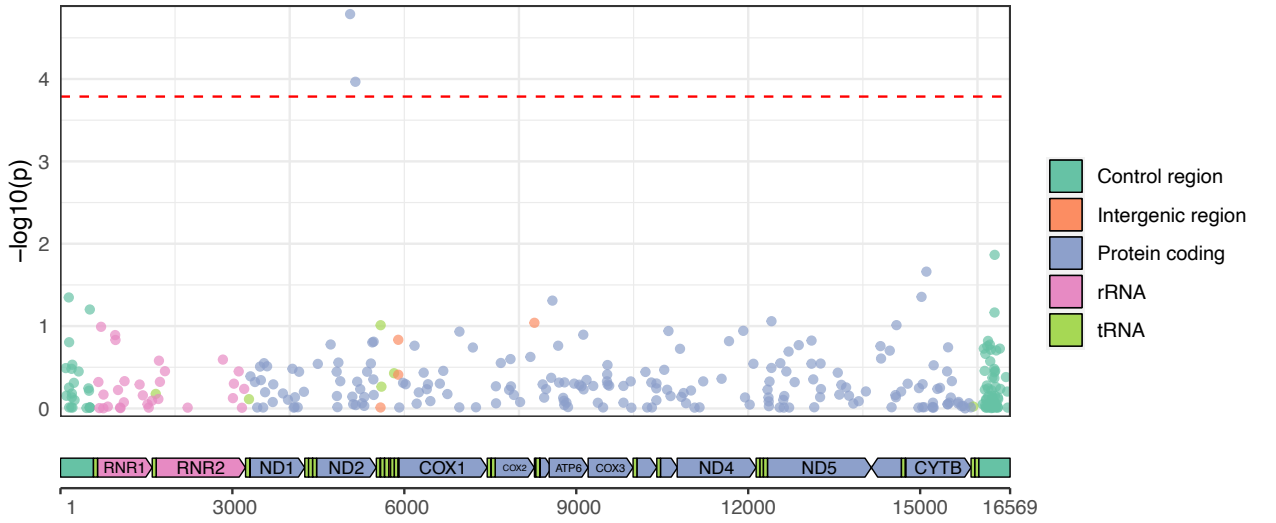


Figure 3. Manhattan plots of mtDNA variant associations with complex traits

The Manhattan plots show the association results of mtDNA variants with multiple complex traits analyzed using a generalized linear model (GLM). Each dot represents a mtDNA variant analyzed for association with a specific phenotype, sorted by position along the x-axis. The colors represent different phenotypes, each denoted by a unique color. Association analyses for quantitative and binary traits are shown in a and b, respectively. Two significance thresholds are shown. The lower red line represents a lenient threshold adjusted by the number of variants ($P = 1.4 \times 10^{-4}$), while the upper orange line represents a stricter threshold adjusted further by the number of phenotypes.



ID	Position	EA	ALT	REF	phenotype	Beta (SE)	OR	p	Gene.Name	Status	INFO	EAF
rs28570593	5054	A	A	G	MYOPIA_1000	0.906 (0.21)	2.474	1.63E-05	*ND2*	imputed	0.938	0.011
rs367778601	5147	A	A	G	MYOPIA_1000	0.671 (0.173)	1.956	1.08E-04	*ND2*	genotyped	1	0.024

Figure 4. Manhattan plot of mtDNA variants associated with high myopia.
The plot illustrates the -log10(p-values) of mtDNA variants analyzed for their association with high myopia, sorted by their position on the mitochondrial genome. Variants are color-coded based on their genomic location: control region (green), intergenic region (orange), protein-coding region (purple), rRNA genes (pink), and tRNA genes (light green). The red dashed line indicates the lenient significance threshold.

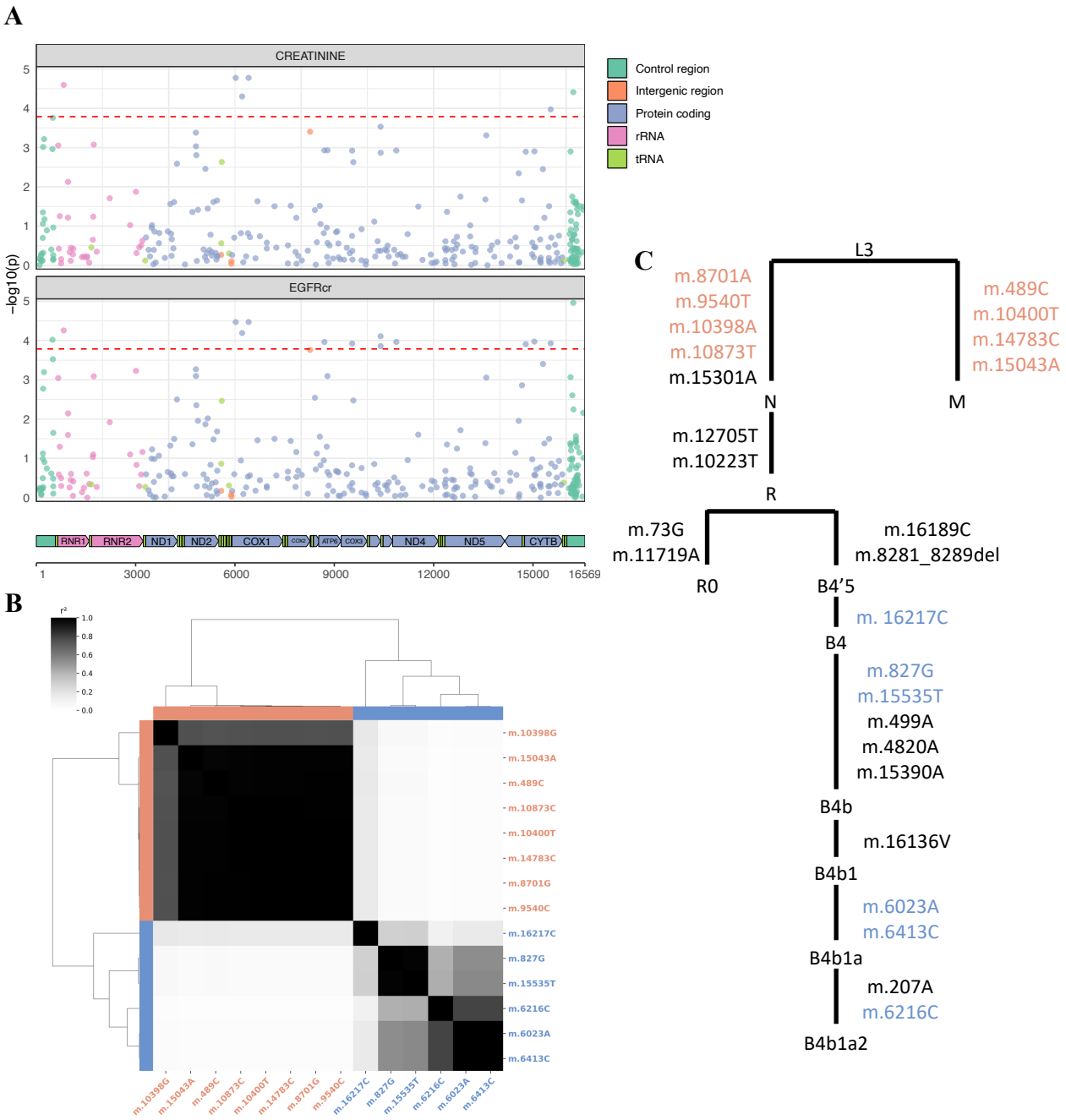


Figure 5. mtDNA variants associated with renal function and phylogenetic lineages

(A) The Manhattan plot illustrates mtDNA variants assessed for their association with renal function biomarkers: Scr levels and eGFR. Variants are color-coded based on their genomic location: control region (green), intergenic region (orange), protein-coding region (purple), rRNA genes (pink), and tRNA genes (light green). The red dashed line indicates the lenient significance threshold.

(B) The heatmap displays the pairwise LD r^2 among the identified mtDNA variants, clustered to highlight two main association sets (orange and blue). LD values were calculated using PLINK.

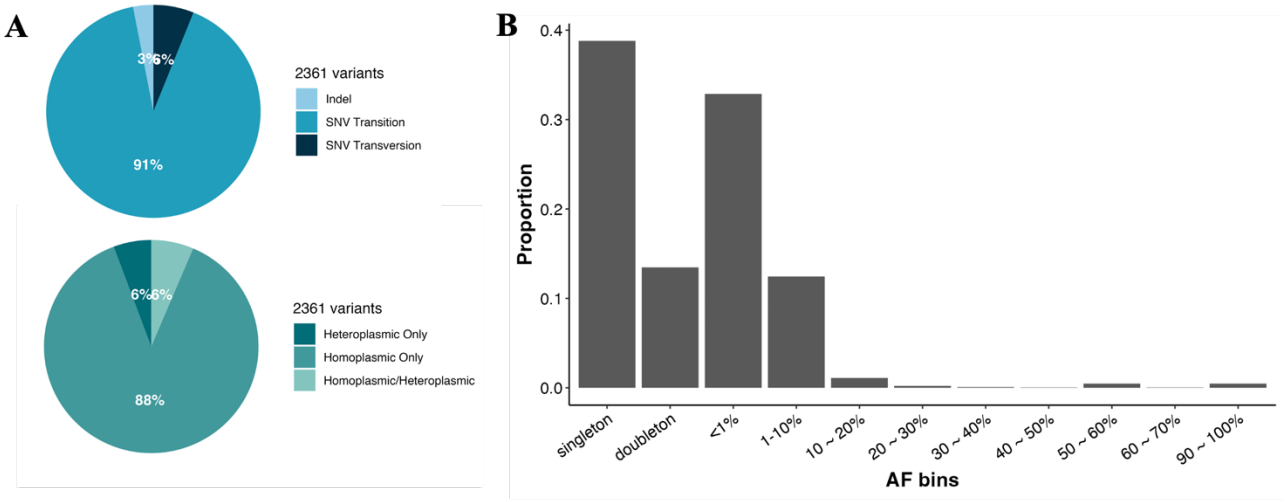


Figure S1. Distribution and classification of mtDNA variants in WGS analyses.
(A) The left pie chart depicts the proportion of variants by different variants. The right pie chart shows the classification of variants that are homoplasmic-only, heteroplasmic-only, or occurring in both types.
(B) The bar graph of AF bins for mtDNA.

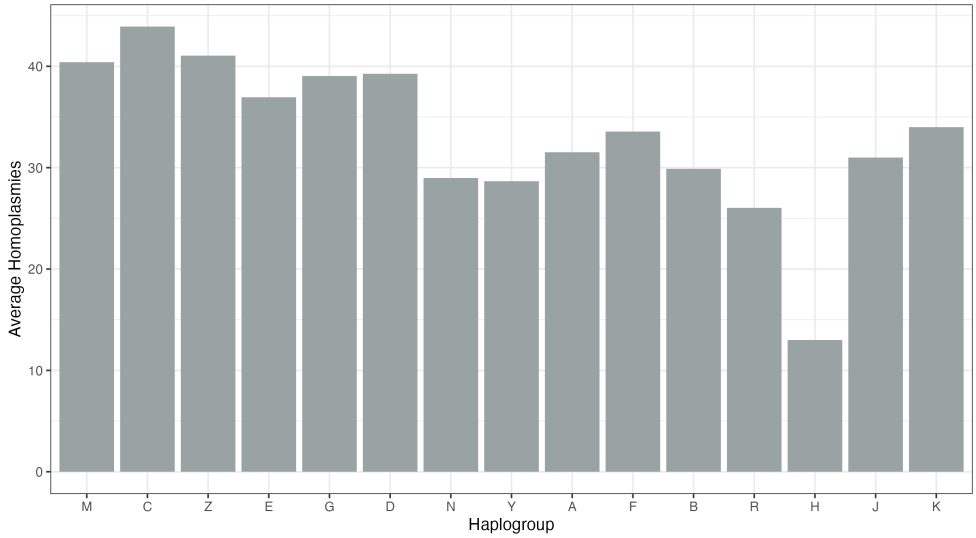


Figure S2. Average homoplasmies within haplogroups
The bar graph displays the average number of mtDNA homoplasmic variants individuals carry within each haplogroup. The order of haplogroup was aligned with Figure 2A.

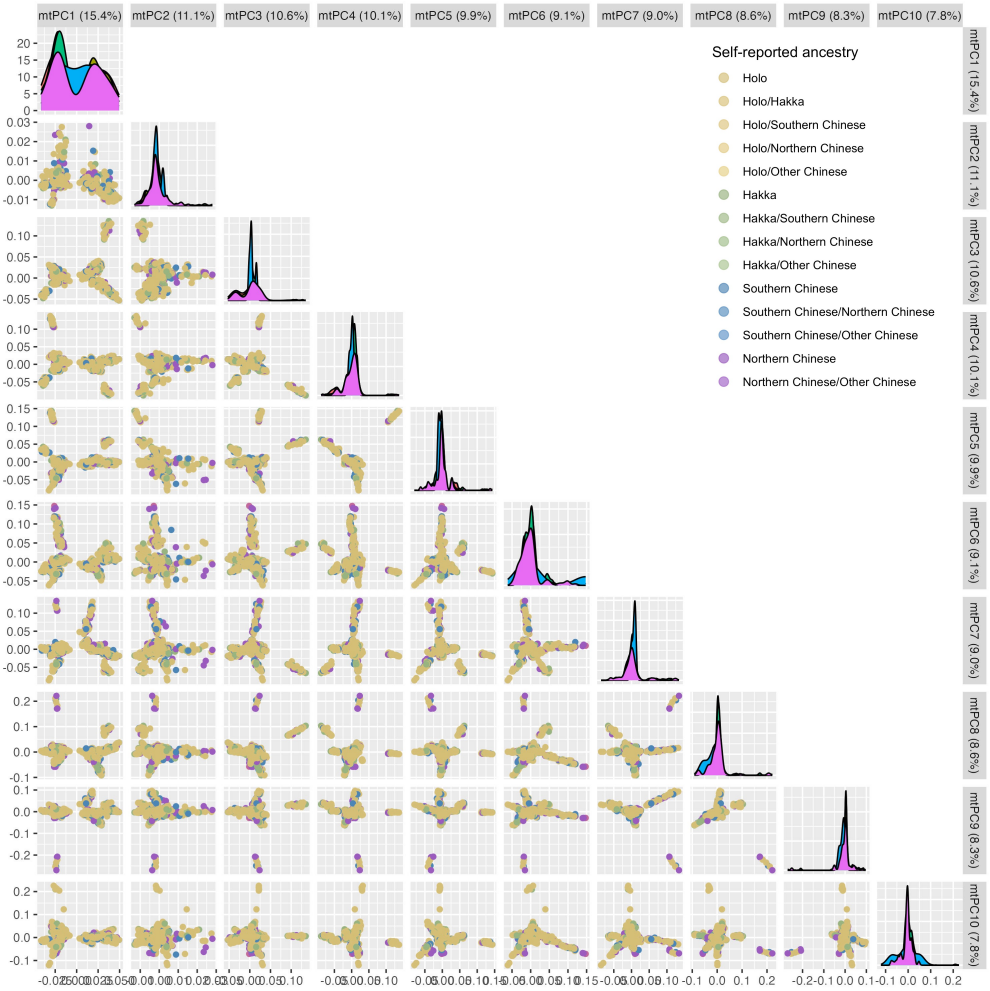
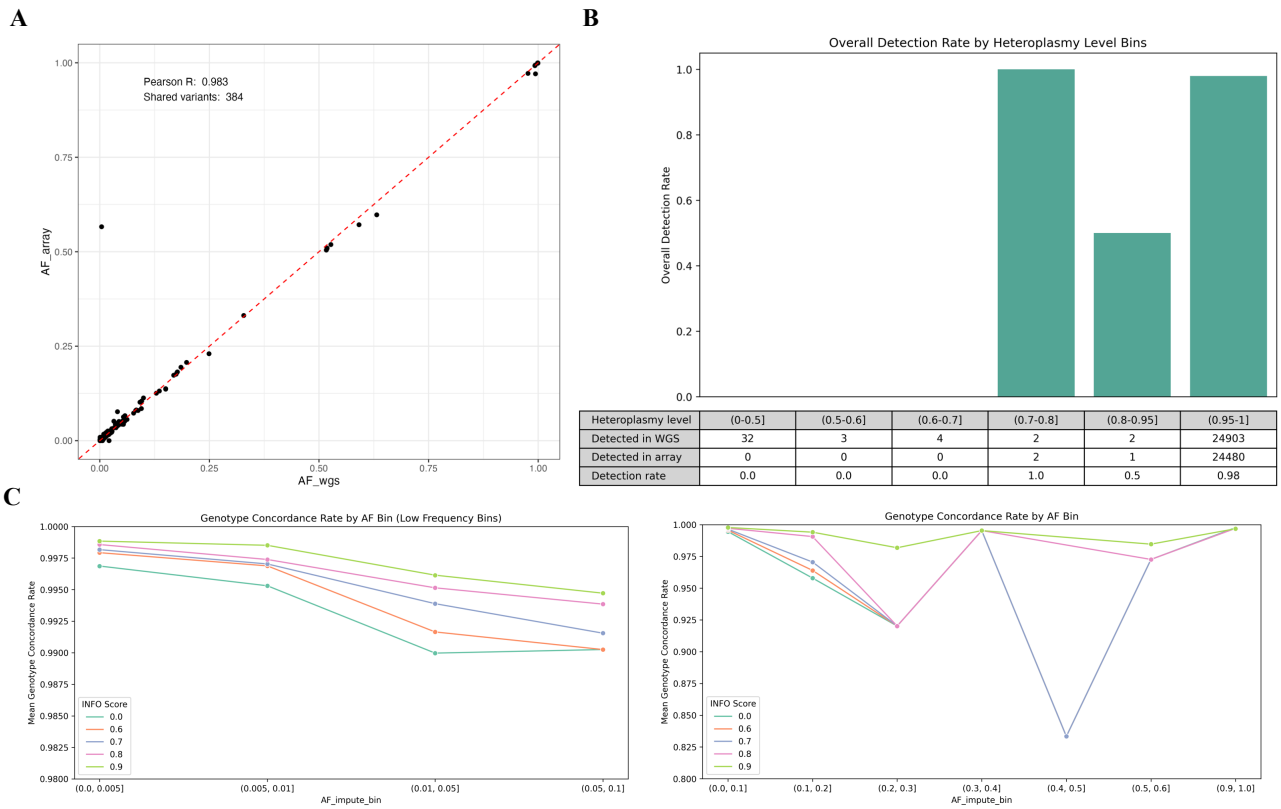


Figure S3. Pairwise plots of mitochondrial PCs (mtPC1 to mtPC10).
Participants are colored according to their self-reported maternal ancestry, with density plots shown along the diagonal.



Heteroplasmy level	(0-0.5]	(0.5-0.6]	(0.6-0.7]	(0.7-0.8]	(0.8-0.95]	(0.95-1]
Detected in WGS	32	3	4	2	2	24903
Detected in array	0	0	0	2	1	24480
Detection rate	0.0	0.0	0.0	1.0	0.5	0.98

Figure S4. Evaluation of mtDNA variants imputation

(A) The plot illustrates the comparison of AFs between microarray and WGS data.

(B) The bar graph displays the overall detection rates of mtDNA variants by heteroplasmy levels. Below the graph, a table details the number of genotype calls detected in both WGS and the microarray, along with the microarray's detection rate.

(C) This plot demonstrates the genotype concordance rates across various AF bins at different INFO score thresholds, with separate analyses for rare and common variants presented in the left and right panels, respectively.

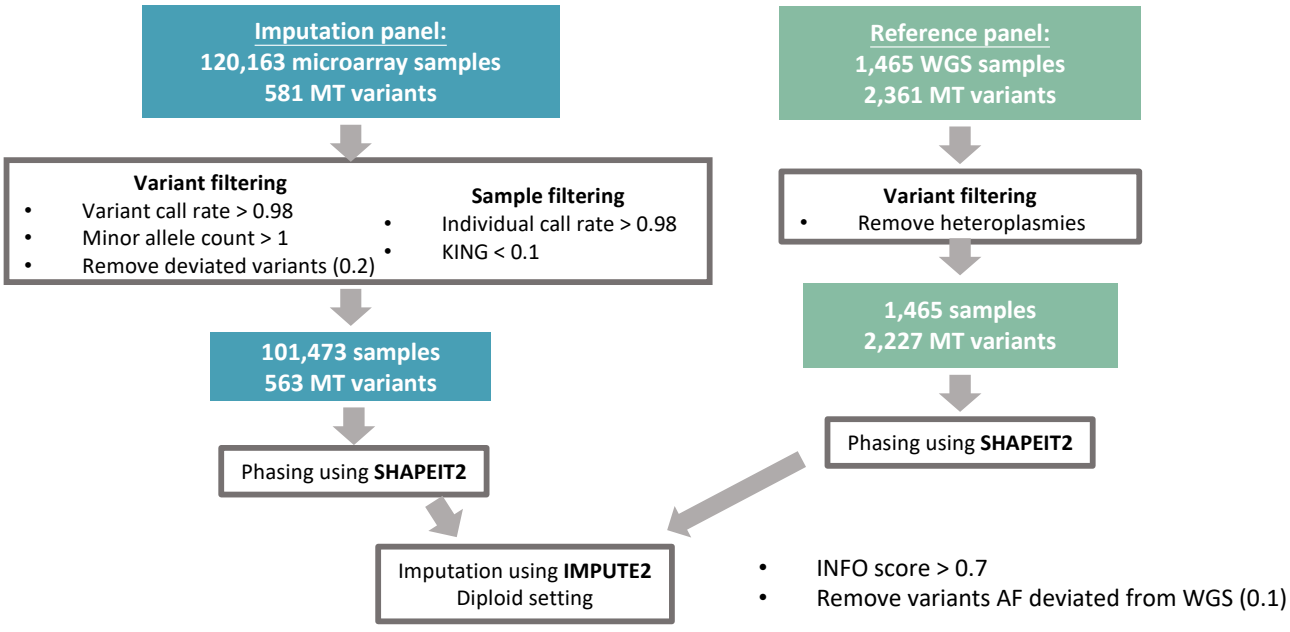


Figure S5. Overview of the mtDNA variant imputation process.

The flowchart outlines the imputation workflow, including participant and variant counts, filtering criteria, and software (bold) used throughout the analysis.

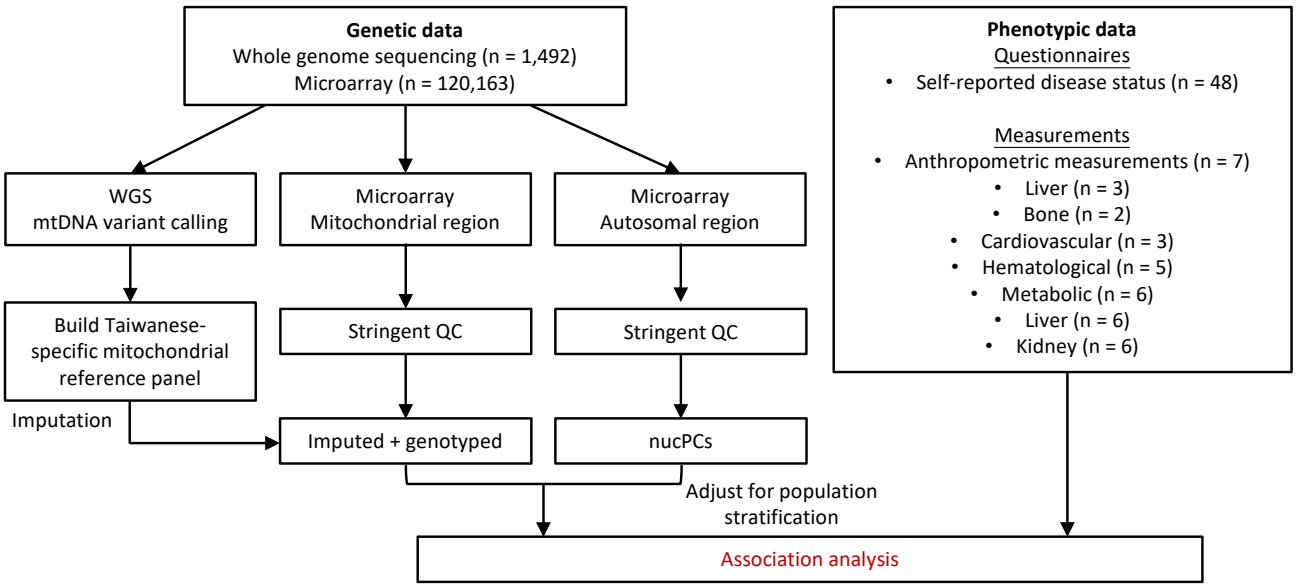


Figure S6. Mitochondrial genome-wide association analysis workflow
The schematic outlines the workflow for mtDNA variant association analysis using genetic data from 1,492 WGS and 120,163 microarray samples. This includes mtDNA variant calling, quality control, and imputation, followed by association analyses with phenotypic data from questionnaires and measurements relevant to complex traits.

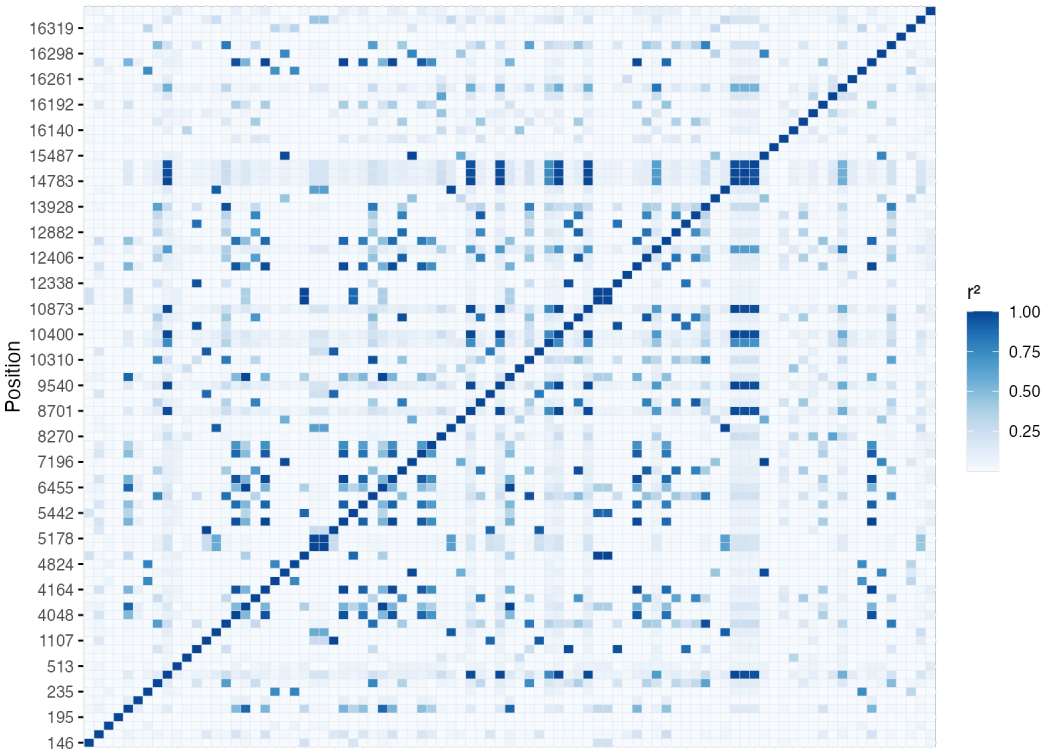


Figure S7. Pairwise LD matrix of mtDNA variants
The heatmap illustrates the pairwise LD patterns across homoplasmic mitochondrial DNA variants with an AF ≥0.05. Each cell represents the r^2 value between pairs of mtDNA variants, indicating the level of correlation. Darker shades indicate higher r^2.

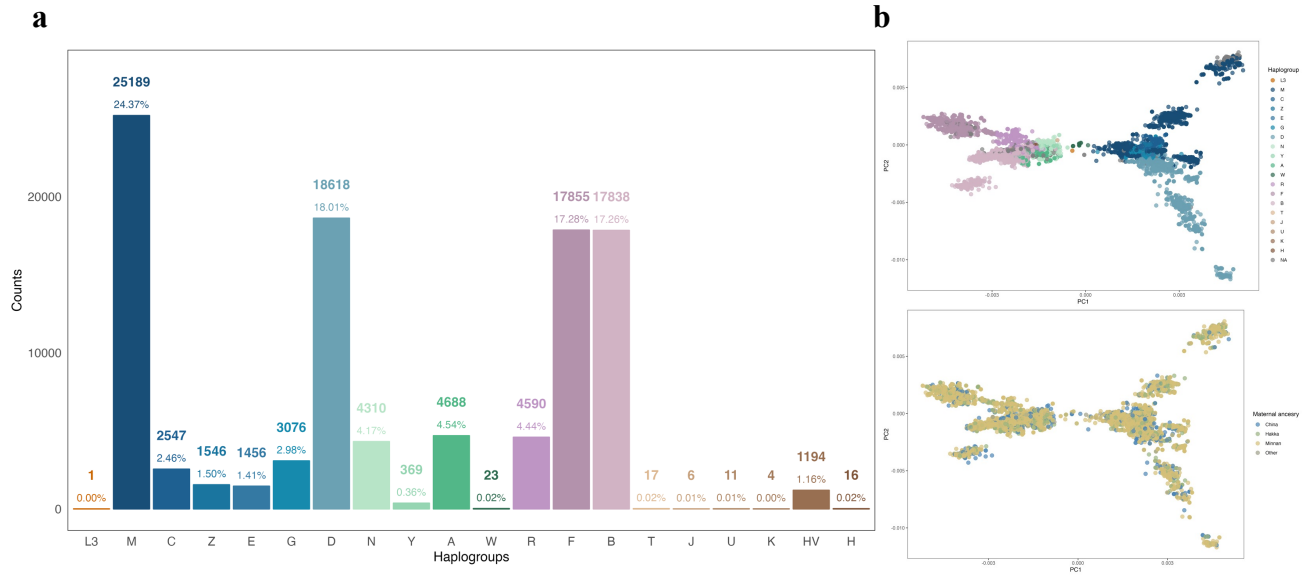


Figure S8. mtDNA haplogroups analysis in TWB using genotyping array

(A) The bar graph shows the prevalence of mtDNA haplogroups among TWB participants, with assignments based on microarray genotypes. The haplogroups are organized according to their relationships in the mitochondrial Phylotree. Haplogroups deriving from macrohaplogroup M are colored in shades of blue, those from N in shades of green, and those from R in shades of purple and brown.

(B and C) PCA analyses of mtDNA variants derived from the genotyping array. The upper plot illustrates individuals colored by their haplogroup assignment, and the lower plot is colored by self-reported maternal ancestry.

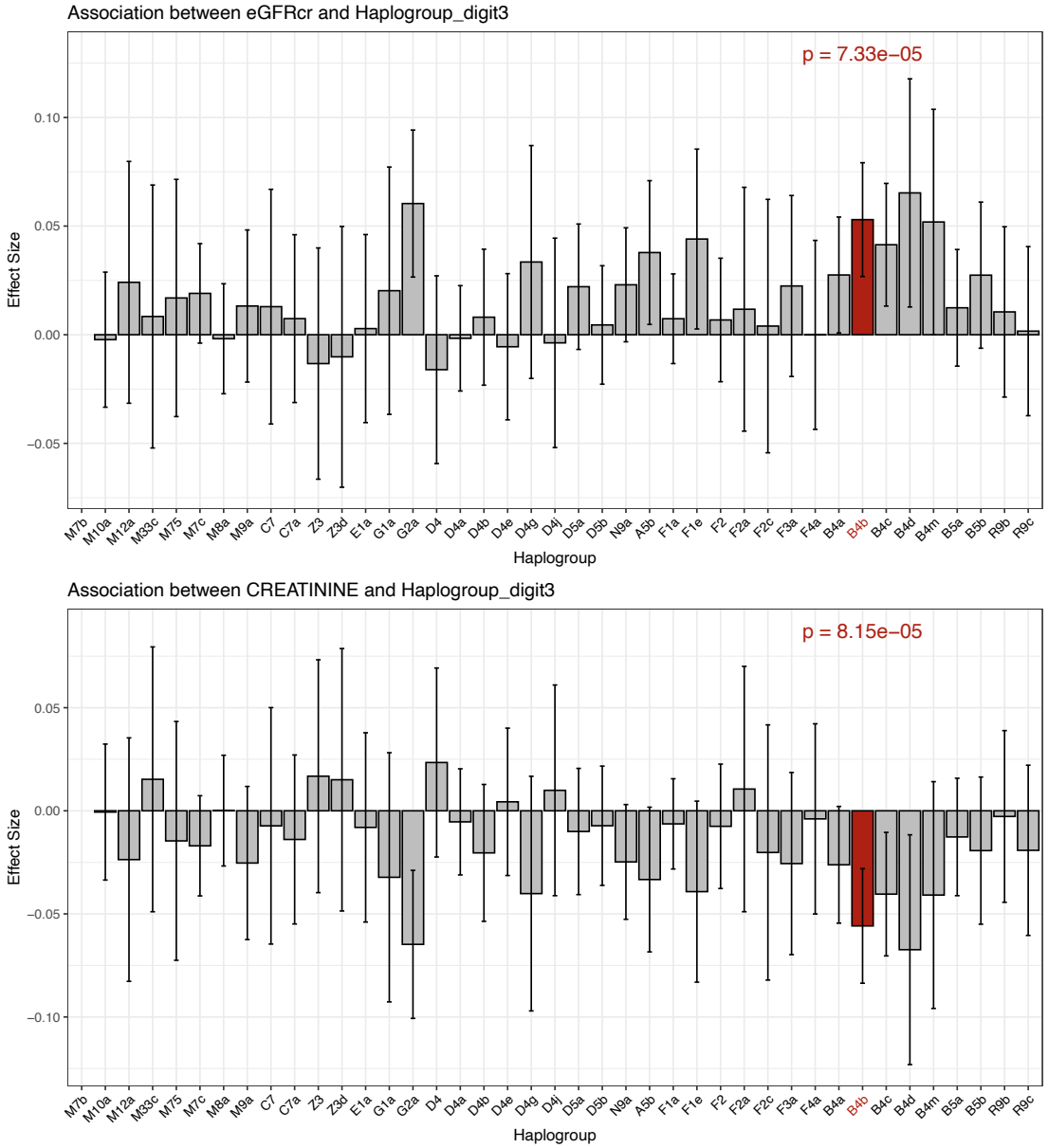


Figure S9. Association of mtDNA haplogroups with renal function markers

The figure displays the associations between mitochondrial haplogroups and two renal function markers: eGFR and creatinine levels. The top panel illustrates the effect sizes of each haplogroup on eGFR, while the bottom panel displays the effect sizes on Scr.

Tables

Table 1. Renal function-associated mtDNA variants in the Taiwanese population.

Phenotype	POS	REF	ALT	Gene (Consequence)	p	Beta (SE)	Status	INFO	Freq.	Haplogroup
Scr	827	A	G	*MT-RNR1* (non_coding_transcript_exon_variant)	2.54E-05	-0.023 (0.006)	genotyped	1	0.044	B4b
Scr	6023	G	A	*MT-CO1* (synonymous_variant)	1.68E-05	-0.032 (0.007)	imputed	0.982	0.024	B4b1a
Scr	6216	T	C	*MT-CO1* (synonymous_variant)	4.99E-05	-0.029 (0.007)	genotyped	1	0.025	B4b1a2
Scr	6413	T	C	*MT-CO1* (synonymous_variant)	1.68E-05	-0.032 (0.007)	imputed	0.995	0.024	B4b1a
Scr	15535	C	T	*MT-CYB* (synonymous_variant)	1.07E-04	-0.021 (0.006)	genotyped	1	0.044	B4b
Scr	16217	T	C	D-LOOP (intergenic_variant)	3.89E-05	-0.014 (0.003)	imputed	0.941	0.135	B4
eGFRcr	489	T	C	D-LOOP (intergenic_variant)	9.53E-05	-0.008 (0.002)	imputed	0.971	0.51	M
eGFRcr	827	A	G	*MT-RNR1* (non_coding_transcript_exon_variant)	5.52E-05	0.021 (0.005)	genotyped	1	0.044	B4b
eGFRcr	6023	G	A	*MT-CO1* (synonymous_variant)	3.37E-05	0.029 (0.007)	imputed	0.982	0.024	B4b1a
eGFRcr	6216	T	C	*MT-CO1* (synonymous_variant)	6.43E-05	0.027 (0.007)	genotyped	1	0.025	B4b1a2
eGFRcr	6413	T	C	*MT-CO1* (synonymous_variant)	3.37E-05	0.029 (0.007)	imputed	0.995	0.024	B4b1a
eGFRcr	8701	A	G	*MT-ATP6* (missense_variant)	1.09E-04	-0.008 (0.002)	imputed	0.995	0.508	N (m.8701A)
eGFRcr	9540	T	C	*MT-CO3* (synonymous_variant)	1.19E-04	-0.008 (0.002)	imputed	0.998	0.508	N (m.9540T)
eGFRcr	10398	A	G	*MT-ND3* (missense_variant)	7.77E-05	-0.008 (0.002)	genotyped	1	0.573	N (m.10398A)
eGFRcr	10400	C	T	*MT-ND3* (synonymous_variant)	1.39E-04	-0.008 (0.002)	genotyped	1	0.508	M
eGFRcr	10873	T	C	*MT-ND4* (synonymous_variant)	1.09E-04	-0.008 (0.002)	genotyped	1	0.505	N (m.10873T)
eGFRcr	14783	T	C	*MT-CYB* (synonymous_variant)	1.24E-04	-0.008 (0.002)	imputed	0.996	0.509	M
eGFRcr	15043	G	A	*MT-CYB* (synonymous_variant)	1.06E-04	-0.008 (0.002)	genotyped	1	0.509	M
eGFRcr	15535	C	T	*MT-CYB* (synonymous_variant)	1.18E-04	0.02 (0.005)	genotyped	1	0.044	B4b
eGFRcr	16217	T	C	D-LOOP (intergenic_variant)	1.09E-05	0.014 (0.003)	imputed	0.941	0.135	B4

The table presents renal function-associated variants, detailing their positions in the reference genome NC_012920.1, reference (REF) and alternative (ALT) alleles, gene location, variant consequences, and defining haplogroups. Statistical results are presented with p-values, effect sizes (Beta), and standard errors (SE) of the association between the ALT allele and the phenotype. Additional information includes imputation status, IMPUTE2 information scores (INFO), and allele frequencies from genotyping arrays in the TWB.

References

1. Saraste, M. (1999). Oxidative phosphorylation at the fin de siècle. Science *283*, 1488–1493. https://doi.org/10.1126/science.283.5407.1488.

2. Gustafsson, C.M., Falkenberg, M., and Larsson, N.-G. (2016). Maintenance and Expression of Mammalian Mitochondrial DNA. Annu. Rev. Biochem. *85*, 133–160. https://doi.org/10.1146/annurev-biochem-060815-014402.

3. Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., et al. (1981). Sequence and organization of the human mitochondrial genome. Nature *290*, 457–465. https://doi.org/10.1038/290457a0.

4. 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. https://doi.org/10.1002/1873-3468.14021.

5. Giles, R.E., Blanc, H., Cann, H.M., and Wallace, D.C. (1980). Maternal inheritance of human mitochondrial DNA. Proc. Natl. Acad. Sci. U. S. A. *77*, 6715–6719. https://doi.org/10.1073/pnas.77.11.6715.

6. Behar, D.M., Rosset, S., Blue-Smith, J., Balanovsky, O., Tzur, S., Comas, D., Mitchell, R.J., Quintana-Murci, L., Tyler-Smith, C., Wells, R.S., et al. (2007). The Genographic Project Public Participation Mitochondrial DNA Database. PLOS Genet. *3*, e104. https://doi.org/10.1371/journal.pgen.0030104.

7. Wallace, D.C. (2015). Mitochondrial DNA variation in human radiation and disease. Cell *163*, 33–38. https://doi.org/10.1016/j.cell.2015.08.067.

8. Poulton, J., Luan, J., Macaulay, V., Hennings, S., Mitchell, J., and Wareham, N.J. (2002). Type 2 diabetes is associated with a common mitochondrial variant: evidence from a population-based case-control study. Hum. Mol. Genet. *11*, 1581–1583. https://doi.org/10.1093/hmg/11.13.1581.

9. Khogali, S.S., Mayosi, B.M., Beattie, J.M., McKenna, W.J., Watkins, H., and Poulton, J. (2001). A common mitochondrial DNA variant associated with susceptibility to dilated cardiomyopathy in two different populations. Lancet Lond. Engl. *357*, 1265–1267. https://doi.org/10.1016/S0140-6736(00)04422-6.

10. Emma, F., Montini, G., Parikh, S.M., and Salviati, L. (2016). Mitochondrial dysfunction in inherited renal disease and acute kidney injury. Nat. Rev. Nephrol. *12*, 267–280. https://doi.org/10.1038/nrneph.2015.214.

11. Borsche, M., Pereira, S.L., Klein, C., and Grünewald, A. (2021). Mitochondria and Parkinson's Disease: Clinical, Molecular, and Translational Aspects. J. Park. Dis. *11*, 45–60. https://doi.org/10.3233/JPD-201981.

12. Yamamoto, K., Sakaue, S., Matsuda, K., Murakami, Y., Kamatani, Y., Ozono, K., Momozawa, Y., and Okada, Y. (2020). Genetic and phenotypic landscape of the mitochondrial genome in the Japanese population. Commun. Biol. *3*, 1–11. https://doi.org/10.1038/s42003-020-0812-9.

13. Yonova-Doing, E., Calabrese, C., Gomez-Duran, A., Schon, K., Wei, W., Karthikeyan, S., Chinnery, P.F., and Howson, J.M.M. (2021). An atlas of mitochondrial DNA genotype–

phenotype associations in the UK Biobank. Nat. Genet. *53*, 982–993. https://doi.org/10.1038/s41588-021-00868-1.

14. Wei, C.-Y., Yang, J.-H., Yeh, E.-C., Tsai, M.-F., Kao, H.-J., Lo, C.-Z., Chang, L.-P., Lin, W.-J., Hsieh, F.-J., Belsare, S., et al. (2021). Genetic profiles of 103,106 individuals in the Taiwan Biobank provide insights into the health and history of Han Chinese. Npj Genomic Med. *6*, 1–10. https://doi.org/10.1038/s41525-021-00178-9.

15. Feng, Y.-C.A., Chen, C.-Y., Chen, T.-T., Kuo, P.-H., Hsu, Y.-H., Yang, H.-I., Chen, W.J., Su, M.-W., Chu, H.-W., Shen, C.-Y., et al. (2022). Taiwan Biobank: A rich biomedical research database of the Taiwanese population. Cell Genomics *2*, 100197. https://doi.org/10.1016/j.xgen.2022.100197.

16. Laricchia, K.M., Lake, N.J., Watts, N.A., Shand, M., Haessly, A., Gauthier, L., Benjamin, D., Banks, E., Soto, J., Garimella, K., et al. (2022). Mitochondrial DNA variation across 56,434 individuals in gnomAD. Genome Res. *32*, 569–582. https://doi.org/10.1101/gr.276013.121.

17. 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*, W58-63. https://doi.org/10.1093/nar/gkw233.

18. Dür, A., Huber, N., and Parson, W. (2021). Fine-Tuning Phylogenetic Alignment and Haplogrouping of mtDNA Sequences. Int. J. Mol. Sci. *22*, 5747. https://doi.org/10.3390/ijms22115747.

19. Lott, M.T., Leipzig, J.N., Derbeneva, O., Xie, H.M., Chalkia, D., Sarmady, M., Procaccio, V., and Wallace, D.C. (2013). mtDNA Variation and Analysis Using Mitomap and Mitomaster. Curr. Protoc. Bioinforma. *44*, 1.23.1-26. https://doi.org/10.1002/0471250953.bi0123s44.

20. Hsu, J.S., Wu, D.-C., Shih, S.-H., Liu, J.-F., Tsai, Y.-C., Lee, T.-L., Chen, W.-A., Tseng, Y.-H., Lo, Y.-C., Lin, H.-Y., et al. (2023). Complete genomic profiles of 1496 Taiwanese reveal curated medical insights. J. Adv. Res. https://doi.org/10.1016/j.jare.2023.12.018.

21. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. *81*, 559–575.

22. Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. Nat. Methods *9*, 179–181. https://doi.org/10.1038/nmeth.1785.

23. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLOS Genet. *5*, e1000529. https://doi.org/10.1371/journal.pgen.1000529.

24. 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., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.

25. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873. https://doi.org/10.1093/bioinformatics/btq559.

26. Ding, J., Sidore, C., Butler, T., Wing, M., Qian, Y., Meirelles, O., Busonero, F., Tsoi, L., Maschio, A., Angius, A., et al. (2015). Assessing Mitochondrial DNA Variation and Copy Number in Lymphocytes of ~2,000 Sardinians Using Tailored Sequencing Analysis Tools. PLoS Genet. *11*, e1005306. https://doi.org/10.1371/journal.pgen.1005306.

27. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation - van Oven - 2009 - Human Mutation - Wiley Online Library https://onlinelibrary.wiley.com/doi/10.1002/humu.20921.

28. Gupta, R., Kanai, M., Durham, T.J., Tsuo, K., McCoy, J.G., Kotrys, A.V., Zhou, W., Chinnery, P.F., Karczewski, K.J., Calvo, S.E., et al. (2023). Nuclear genetic control of mtDNA copy number and heteroplasmy in humans. Nature *620*, 839–848. https://doi.org/10.1038/s41586-023-06426-5.

29. Gorman, G.S., Chinnery, P.F., DiMauro, S., Hirano, M., Koga, Y., McFarland, R., Suomalainen, A., Thorburn, D.R., Zeviani, M., and Turnbull, D.M. (2016). Mitochondrial diseases. Nat. Rev. Dis. Primer *2*, 1–22. https://doi.org/10.1038/nrdp.2016.80.

30. Xing, S., Jiang, S., Wang, S., Lin, P., Sun, H., Peng, H., Yang, J., Kong, H., Wang, S., Bai, Q., et al. (2023). Association of mitochondrial DNA variation with high myopia in a Han Chinese population. Mol. Genet. Genomics *298*, 1059–1071. https://doi.org/10.1007/s00438-023-02036-y.

31. Elson, J.L., Andrews, R.M., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., and Howell, N. (2001). Analysis of European mtDNAs for Recombination. Am. J. Hum. Genet. *68*, 145–153.

32. Pfanner, N., Warscheid, B., and Wiedemann, N. (2019). Mitochondrial proteins: from biogenesis to functional networks. Nat. Rev. Mol. Cell Biol. *20*, 267–284. https://doi.org/10.1038/s41580-018-0092-0.

33. Brown, M.D., Torroni, A., Reckord, C.L., and Wallace, D.C. (1995). Phylogenetic analysis of Leber's hereditary optic neuropathy mitochondrial DNA's indicates multiple independent occurrences of the common mutations. Hum. Mutat. *6*, 311–325. https://doi.org/10.1002/humu.1380060405.

34. Brown, M.D., Zhadanov, S., Allen, J.C., Hosseini, S., Newman, N.J., Atamonov, V.V., Mikhailovskaya, I.E., Sukernik, R.I., and Wallace, D.C. (2001). Novel mtDNA mutations and oxidative phosphorylation dysfunction in Russian LHON families. Hum. Genet. *109*, 33–39. https://doi.org/10.1007/s004390100538.

35. Govers, L.P., Toka, H.R., Hariri, A., Walsh, S.B., and Bockenhauer, D. (2021). Mitochondrial DNA mutations in renal disease: an overview. Pediatr. Nephrol. *36*, 9–17. https://doi.org/10.1007/s00467-019-04404-6.

36. Cañadas-Garre, M., Baños-Jaime, B., Maqueda, J.J., Smyth, L.J., Cappa, R., Skelly, R., Hill, C., Brennan, E.P., Doyle, R., Godson, C., et al. (2024). Genetic variants affecting mitochondrial function provide further insights for kidney disease. BMC Genomics *25*, 576. https://doi.org/10.1186/s12864-024-10449-1.

37. Trejaut, J.A., Kivisild, T., Loo, J.H., Lee, C.L., He, C.L., Hsu, C.J., Li, Z.Y., and Lin, M. (2005). Traces of Archaic Mitochondrial Lineages Persist in Austronesian-Speaking Formosan Populations. PLoS Biol. *3*, e247. https://doi.org/10.1371/journal.pbio.0030247.

38. Cai, N., Gomez-Duran, A., Yonova-Doing, E., Kundu, K., Burgess, A.I., Golder, Z.J., Calabrese, C., Bonder, M.J., Camacho, M., Lawson, R.A., et al. (2021). Mitochondrial DNA variants modulate N-formylmethionine, proteostasis and risk of late-onset human diseases. Nat. Med. *27*, 1564–1575. https://doi.org/10.1038/s41591-021-01441-3.

39. Stewart, J.B., and Chinnery, P.F. (2015). The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. Nat. Rev. Genet. *16*, 530–542. https://doi.org/10.1038/nrg3966.

40. Morava, E., Kozicz, T., and Wallace, D.C. (2019). The phenotype modifier: is the mitochondrial DNA background responsible for individual differences in disease severity. J. Inherit. Metab. Dis. *42*, 3–4. https://doi.org/10.1002/jimd.12050.