

2023

Association analysis and clustering of rare variants with disease phenotypes

<https://hdl.handle.net/2144/52801>

"Downloaded from OpenBU. Boston University's institutional repository."

BOSTON UNIVERSITY
GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

**ASSOCIATION ANALYSIS AND CLUSTERING
OF RARE VARIANTS WITH DISEASE PHENOTYPES**

by

XIANBANG SUN

B.S., Stony Brook University, 2013
M.S., University of Washington, 2015

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2023

Approved by

First Reader

Chunyu Liu, Ph.D.
Associate Professor of Biostatistics

Second Reader

Josée Dupuis, Ph.D.
Professor of Biostatistics

Third Reader

Kathryn L. Lunetta, Ph.D.
Professor of Biostatistics

Fourth Reader

Seung Hoan Choi, Ph.D.
Research Assistant Professor of Biostatistics

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my esteemed dissertation advisor, Dr. Chunyu Liu, for her invaluable supervision, support, and tutelage during the course of my Ph.D. degree. This work would not have been possible without her consistent and strong support. My gratitude extends to my committee members, Dr. Josée Dupuis, Dr. Kathryn L. Lunetta, and Dr. Seung Hoan Choi, for their treasured support and encouragement. Additionally, I would like to thank Katia Bulekova and Dr. Achilleas N. Pitsillides for their valuable help.

Finally, I would like to express my very profound gratitude to my parents, my wife, and my little son for providing me with unfailing support and continuous encouragement throughout my years of Ph.D. study, the process of research, and writing this thesis. This accomplishment would not have been possible without them. Thank you.

**ASSOCIATION ANALYSIS AND CLUSTERING
OF RARE VARIANTS WITH DISEASE PHENOTYPES**

XIANBANG SUN

Boston University Graduate School of Arts and Sciences, 2023

Primary Mentor: Chunyu Liu, Associate Professor of Biostatistics

ABSTRACT

Hundreds of thousands of human deoxyribonucleic acid (DNA) samples have been whole genome sequenced, identifying numerous rare variants in the nuclear genome and mitochondrial genome (mtDNA). Multiple mtDNA molecules are present in a cell. mtDNA heteroplasmy is the presence of two or more nucleotides at an mtDNA location in the same individual. Most of the heteroplasmic variants are extremely rare, posing a challenge to applying traditional analytic approaches in association with heteroplasmy. On the other hand, clustering disease-associated rare variants (e.g., classify them into null, positively, or negatively associated groups) in a gene region provides useful information for investigating the underlying biological mechanisms between rare variants and disease traits. However, few studies have investigated rare variants clustering.

To fill in these knowledge gaps, this dissertation focuses on association analysis and clustering of rare variants. In project 1, we develop and evaluate a comprehensive framework for association testing of heteroplasmy using both simulated and real data. In project 2, we propose a method to cluster trait-

associated rare variants based on a Gaussian mixture model (GMM) and apply this method to a real dataset. We also assess the effect of linkage disequilibrium (LD) on the performance of the clustering method in simulation studies. In project 3, we apply the framework developed in project 1 for association analysis of heteroplasmy to cardiometabolic diseases (CMDs) in six TOPMed cohorts to identify CMD-associated heteroplasmic gene regions. Knowledge gained from these three projects will help to better understand the role of rare genetic variants in the etiology of complex human diseases.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	iv
ABSTRACT	v
TABLE OF CONTENTS	vii
LIST OF TABLES	x
LIST OF FIGURES	xx
LIST OF ABBREVIATIONS	xxiii
CHAPTER 1 INTRODUCTION	1
1.1 Introduction.....	1
1.2 Nuclear Inheritance and Mitochondrial Inheritance	2
1.3 Mitochondrial DNA Reference Sequences	3
1.4 mtDNA Homoplasmy and Heteroplasmy	4
1.5 Association Analysis of mtDNA Variants	6
1.6 Pipelines for the Identification of mtDNA Sequencing Variations	7
1.7 Gene-based Association Tests for Nuclear DNA.....	8
1.8 Outline of Dissertation	10
CHAPTER 2 ASSOCIATION ANALYSIS OF MITOCHONDRIAL DNA HETEROPLASMIC VARIANTS: METHODS AND APPLICATIONS.....	14
2.1 Introduction.....	14
2.2 Methods.....	16
2.2.1 Definition of mtDNA Sequence Variants	16
2.2.2 A Phenotype Model.....	17
2.2.3 Gene-based Tests	18
2.2.4 Combing Multiple Gene-based Tests in a Gene/Region	24
2.2.5 A Simulation Study.....	25
2.2.6 Application of the Proposed Framework in Real Data.....	29
2.3 Results	32

2.3.1 Empirical Type I Error Rate of Simulation Studies.....	32
2.3.2 Empirical Statistical Power of Simulation Studies.....	34
2.3.3 Application to Real Data	37
2.4 Discussion.....	40
2.5 Tables and Figures.....	45
CHAPTER 3 CLUSTERING OF RARE VARIANTS FOR CAUSAL VARIANTS IDENTIFICATION AND EFFECT DIRECTION CLASSIFICATION.....	
3.1 Introduction.....	49
3.2 Methods.....	51
3.2.1 Association Testing of Rare Variants	51
3.2.2 Multiple-variant Model to Obtain Variant Level Summary Statistics	52
3.2.3 Gaussian Mixture Model	53
3.2.4 Parameter Estimation by Expectation Maximization (EM) Algorithm ..	55
3.2.5 A Simulation Study.....	59
3.2.6 Application to Exome-wide Association and a Rare-variant GWAS of Blood Pressure Traits	63
3.3 Results	64
3.3.1 Simulation Studies	64
3.3.2 Application to GWAS of BP Traits with Rare Variants.....	68
3.4 Discussion.....	71
3.5 Tables	75
CHAPTER 4 ASSOCIATION ANALYSIS OF MITOCHONDRIAL HETEROPLASMIC VARIANTS AND CARDIOMETABOLIC TRAITS	
4.1 Introduction.....	79
4.2 Methods.....	81
4.2.1 Study Participants	81
4.2.2 Identification of Heteroplasmy.....	81
4.2.3 Cardiometabolic Traits	83

4.2.4 Association Analyses of Rare Heteroplasmic Variants with Cardiometabolic Traits by Gene-based Tests	84
4.3 Results	86
4.3.1 Participants Characteristics.....	86
4.3.2 Heteroplasmy Distribution	87
4.3.3 Association Between Heteroplasmic Burden and Year of Examination	88
4.3.4 Association between Heteroplasmy and CMD Traits	88
4.4 Discussion	90
4.5 Tables	93
CHAPTER 5 SUMMARY AND FUTURE WORK	98
5.1 Summary	98
5.2 Future Work.....	101
5.2.1 mtDNA Genotype Simulation	101
5.2.2 Association Analysis of Heteroplasmy using Correlated Data	102
5.2.3 Accounting for LD Structure Using Summary Statistics in Rare Variants Clustering.....	102
APPENDIX A: SUPPLEMENTARY MATERIALS FOR CHAPTER 2.....	104
A.1 Supplementary Tables	104
A.2 Supplementary Figures	132
APPENDIX B: SUPPLEMENTARY MATERIALS FOR CHAPTER 3.....	138
B.1 Supplementary Tables	138
B.2 Supplementary Figures.....	166
APPENDIX C: SUPPLEMENTARY MATERIALS FOR CHAPTER 4	179
REFERENCES	195
CURRICULUM VITAE	205

LIST OF TABLES

Table 2.1 Gene-wide empirical type I error rates with 95% confidence interval for coding definition 1 in simulation studies at $\alpha=0.001$	45
Table 2.2 Participant characteristics in the five population-level cohorts with whole genome sequencing	46
Table 2.3 Genes showing significant associations with age in the meta-analysis using Fisher's method	47
Table 3.1 Summary of six simulation scenarios with 1000 replicates	75
Table 3.2 MSE of estimated clusters' means for a continuous trait with the absence of LD in simulation studies	76
Table 3.3 Summary of clustering results for rare variants within the combined signal regions of BP traits.....	77
Table 3.4 Summary of clustering results for rare variants within the signal genes of SBP	78
Table 4.1 Cohort-specific characteristics for AA (A.) and EA (B.).....	93
Table 4.2 Cohort-specific distribution of heteroplasmic variants across sixteen mtDNA genes for AA (A.) and EA (B.).....	94
Table 4.3 Associations of heteroplasmic burden with the year of blood draw	96

Table 4.4 Significant associations between heteroplasmic variants and CMD traits across sixteen mtDNA genes by definition 3 for continuous traits (A.) and binary traits (B.).....	97
Supplementary Table 2.1 Frequency of heteroplasmic sites in the CYB gene in simulation studies.....	104
Supplementary Table 2.2 Gene-wise empirical type I error rates with 95% confidence interval for coding definition 2 using simulation data at $\alpha=0.001$	105
Supplementary Table 2.3 Association analysis of heteroplasmic burden with the year of blood draw.....	106
Supplementary Table 2.4 Distribution of heteroplasmic variants in the five cohorts	107
Supplementary Table 2.5 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 1 from Fisher's method meta-analysis for participants of AA ancestry.....	108
Supplementary Table 2.6 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 2 from Fisher's method meta-analysis for participants of AA ancestry.....	109
Supplementary Table 2.7 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 1 from the fixed-effect meta-analysis for participants of AA ancestry	110

Supplementary Table 2.8 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 2 from the fixed-effect meta-analysis for participants of AA ancestry	111
Supplementary Table 2.9 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 1 from Fisher's method meta-analysis for participants of EA ancestry.....	112
Supplementary Table 2.10 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 2 from Fisher's method meta-analysis for participants of EA ancestry.....	113
Supplementary Table 2.11 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 1 from the fixed-effect meta-analysis for participants of EA ancestry	114
Supplementary Table 2.12 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 2 from the fixed-effect meta-analysis for participants of EA ancestry	115
Supplementary Table 2.13 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 1 from Fisher's method meta-analysis for all participants	116

Supplementary Table 2.14 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 2 from Fisher’s method meta-analysis for all participants	117
Supplementary Table 2.15 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 1 from the fixed-effect meta-analysis of all participants	118
Supplementary Table 2.16 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 2 from the fixed-effect meta-analysis of all participants	119
Supplementary Table 2.17 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 1 from Fisher’s method meta-analysis for participants of AA ancestry.....	120
Supplementary Table 2.18 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 2 from Fisher’s method meta-analysis for participants of AA ancestry.....	121
Supplementary Table 2.19 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 1 from the fixed-effect meta-analysis for participants of AA ancestry	122

Supplementary Table 2.20 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 2 from the fixed-effect meta-analysis for participants of AA ancestry	123
Supplementary Table 2.21 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 1 from Fisher’s method meta-analysis for participants of EA ancestry.....	124
Supplementary Table 2.22 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 2 from Fisher’s method meta-analysis for participants of EA ancestry.....	125
Supplementary Table 2.23 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 1 from the fixed-effect meta-analysis for participants of EA ancestry	126
Supplementary Table 2.24 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 2 from the fixed-effect meta-analysis for participants of EA ancestry	127
Supplementary Table 2.25 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 1 from Fisher’s method meta-analysis for all participants	128

Supplementary Table 2.26 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 2 from Fisher’s method meta-analysis for all participants	129
Supplementary Table 2.27 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 1 from the fixed-effect meta-analysis of all participants	130
Supplementary Table 2.28 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 2 from the fixed-effect meta-analysis of all participants	131
Supplementary Table 3.1 MSE of estimated clusters' means for a continuous trait with the presence of LD in simulation studies	138
Supplementary Table 3.2 Accuracy of number of clusters specification for a continuous trait with the absence of LD in simulation studies	139
Supplementary Table 3.3 Accuracy of number of clusters specification for a continuous trait with the presence of LD in simulation studies	140
Supplementary Table 3.4 Accuracy of number of clusters specification for a binary trait with the absence of LD in simulation studies	141
Supplementary Table 3.5 Accuracy of number of clusters specification for a binary trait with the presence of LD in simulation studies	142

Supplementary Table 3.6 Summary of associations between genes and blood pressure traits in real data analysis	143
Supplementary Table 3.7 Summary of clustering results for rare variants within signal genes of DBP.....	144
Supplementary Table 3.8 Summary of clustering results for rare variants within signal genes of PP	145
Supplementary Table 3.9 Summary of clustering results for rare variants within signal genes of HTN.....	146
Supplementary Table 3.10 Clustering results of rare variants within the combined signal region of SBP	147
Supplementary Table 3.11 Clustering results of rare variants within the combined signal region of DBP.....	149
Supplementary Table 3.12 Clustering results of rare variants within the combined signal region of PP	151
Supplementary Table 3.13 Clustering results of rare variants within the combined signal region of HTN.....	153
Supplementary Table 3.14 Clustering results of rare variants within the DBH gene with SBP.....	155
Supplementary Table 3.15 Clustering results of rare variants within the NPR1 gene with SBP.....	156

Supplementary Table 3.16 Clustering results of rare variants within the PLCB3 gene with SBP	157
Supplementary Table 3.17 Clustering results of rare variants within the DBH gene with DBP.....	158
Supplementary Table 3.18 Clustering results of rare variants within the PLCB3 gene with DBP.....	159
Supplementary Table 3.19 Clustering results of rare variants within the CEP120 gene with PP	160
Supplementary Table 3.20 Clustering results of rare variants within the COL21A1 gene with PP	161
Supplementary Table 3.21 Clustering results of rare variants within the NOX4 gene with PP	162
Supplementary Table 3.22 Clustering results of rare variants within the DBH gene with HTN.....	163
Supplementary Table 3.23 Clustering results of rare variants within the NPR1 gene with HTN	164
Supplementary Table 3.24 Clustering results of rare variants within the PLCB3 gene with HTN	165
Supplementary Table 4.1 Associations between heteroplasmic variants and BMI across sixteen mtDNA genes among AA	179

Supplementary Table 4.2 Associations between heteroplasmic variants and BMI across sixteen mtDNA Genes among EA.....	180
Supplementary Table 4.3 Associations between heteroplasmic variants and obesity across sixteen mtDNA genes among AA	181
Supplementary Table 4.4 Associations between heteroplasmic variants and obesity across sixteen mtDNA genes among EA	182
Supplementary Table 4.5 Associations between heteroplasmic variants and SBP across sixteen mtDNA genes among AA	183
Supplementary Table 4.6 Associations between heteroplasmic variants and SBP across sixteen mtDNA genes among EA	184
Supplementary Table 4.7 Associations between heteroplasmic variants and HTN across sixteen mtDNA genes among AA	185
Supplementary Table 4.8 Associations between heteroplasmic variants and HTN across sixteen mtDNA genes among EA	186
Supplementary Table 4.9 Associations between heteroplasmic variants and BG across sixteen mtDNA genes among AA	187
Supplementary Table 4.10 Associations between heteroplasmic variants and BG across sixteen mtDNA genes among EA	188
Supplementary Table 4.11 Associations between heteroplasmic variants and diabetes across sixteen mtDNA genes among AA	189

Supplementary Table 4.12 Associations between heteroplasmic variants and diabetes across sixteen mtDNA genes among EA	190
Supplementary Table 4.13 Associations between heteroplasmic variants and LDL across sixteen mtDNA genes among AA.....	191
Supplementary Table 4.14 Associations between heteroplasmic variants and LDL across sixteen mtDNA genes among EA.....	192
Supplementary Table 4.15 Associations between heteroplasmic variants and hyperlipidemia across sixteen mtDNA genes among AA	193
Supplementary Table 4.16 Associations between heteroplasmic variants and hyperlipidemia across sixteen mtDNA genes among EA	194

LIST OF FIGURES

Figure 2.1 Simulation-based power comparisons of six gene-based tests and two omnibus tests with a continuous and a binary trait for coding definition 1 (adjusted for empirical type I error rate)	48
Supplementary Figure 2.1 Simulation study: power comparisons of six aggregate unit tests and two omnibus tests with a continuous trait and a binary trait for coding definition 2 (adjusted for empirical type I error rate).....	132
Supplementary Figure 2.2 Simulation-based power comparisons of six aggregate unit tests and two omnibus tests with a continuous and a binary trait for coding definition 1 (unadjusted for empirical type I error rate)	133
Supplementary Figure 2.3 Simulation study: power comparisons of six aggregate unit tests and two omnibus tests with a continuous trait and a binary trait for coding definition 2 (unadjusted for empirical type I error rate).....	134
Supplementary Figure 2.4 The proportion of heteroplasmic variants in each of the sixteen genes and D-loop region	135
Supplementary Figure 2.5 Comparison of standardized beta coefficients of simple burden test of coding definition 1 with/o cell count variables in Framingham Heart Study (FHS) and Jackson Heart Study (JHS).....	136
Supplementary Figure 2.6 Simulation study: power comparisons at beta (MAF_H , 1, 1) and beta (MAF_H , 1, 25).....	137

Supplementary Figure 3.1: Heatmap of the average correlation matrix of the 60 simulated rare variants with the presence of LD structure..... 166

Supplementary Figure 3.2: Boxplots of ARI values of simulation scenarios 1, 2, and 3 with combined weighting scheme with/without LD for a continuous trait 167

Supplementary Figure 3.3: Boxplots of ARI values of simulation scenarios 4, 5, and 6 with combined weighting scheme with/without LD for a continuous trait 168

Supplementary Figure 3.4: Boxplots of ARI values of simulation scenarios 1, 2, and 3 with combined weighting scheme with/without LD for a binary trait..... 169

Supplementary Figure 3.5: Boxplots of ARI values of simulation scenarios 4, 5, and 6 with combined weighting scheme with/without LD for a binary trait..... 170

Supplementary Figure 3.6: Boxplots of ARI values of simulation scenarios 1, 2, and 3 with equal, inverse SE and combined weighting schemes with the absence of LD for a continuous trait..... 171

Supplementary Figure 3.7: Boxplots of ARI values of simulation scenarios 4, 5, and 6 with equal, inverse SE and combined weighting schemes with the absence of LD for a continuous trait..... 172

Supplementary Figure 3.8: Boxplots of ARI values of simulation scenarios 1, 2, and 3 with equal, inverse SE and combined weighting schemes with the absence of LD for a binary trait	173
Supplementary Figure 3.9: Boxplots of ARI values of simulation scenarios 4, 5, and 6 with equal, inverse SE and combined weighting schemes with the absence of LD for a binary trait	174
Supplementary Figure 3.10: Boxplots of ARI values of simulation scenarios 1, 2, and 3 with equal, inverse SE and combined weighting schemes with the presence of LD for a continuous trait.....	175
Supplementary Figure 3.11: Boxplots of ARI values of simulation scenarios 4, 5, and 6 with equal, inverse SE and combined weighting schemes with the presence of LD for a continuous trait.....	176
Supplementary Figure 3.12: Boxplots of ARI values of simulation scenarios 1, 2, and 3 with equal, inverse SE and combined weighting schemes with the presence of LD for a binary trait	177
Supplementary Figure 3.13: Boxplots of ARI values of simulation scenarios 4, 5, and 6 with equal, inverse SE and combined weighting schemes with the presence of LD for a binary trait	178

LIST OF ABBREVIATIONS

AA	African Americans
ACAT	aggregated Cauchy association test
ACAT-O	aggregated Cauchy association optimal unified test
ANOVA	Analysis of variance
ARI	Adjusted rand index
ATP	Adenosine triphosphate
BG	Blood glucose
BIC	Bayesian information criterion
BMI	Body mass index
CADD	Combined Annotation Dependent Depletion
CEP120	Centrosomal protein 120 gene
CHARGE	The Cohorts for Heart and Aging Research in Genomic Epidemiology
CHD	Coronary heart disease
CI	Confidence interval
COL21A1	Collagen Type XXI Alpha 1 Chain
DBH	Dopamine beta-hydroxylase
DBP	Diastolic blood pressure
DNA	Deoxyribonucleic acid
EA	European Americans
EM	Expectation–maximization algorithm
ExomeBP	Exome Blood Pressure Consortium
GLM	Generalized linear model
GoT2D	The Genetics of Type 2 Diabetes Consortium
HDL	High-density lipoprotein
HTN	Hypertension
LD	Lineage disequilibrium
LDL	Low-density lipoprotein
LMM	Linear mixed model

MAC	Minor allele count
MAF	Minor allele frequency
MLE	Maximum likelihood estimation
MSE	mean squared error
mtDNA	Mitochondrial DNA
nDNA	Nuclear DNA
NHLBI	National Heart, Lung, and Blood Institute
NOX4	NADPH Oxidase 4
NPR1	Natriuretic Peptide Receptor 1
OXPPOS	Oxidative phosphorylation
<i>PLCB3</i>	Phosphatidylinositol-4,5-bisphosphate phosphodiesterase beta-3
PP	Pulse pressure
rCRS	Revised Cambridge Reference Sequence
rRNA	ribosomal ribonucleic acid
RSRS	Reconstructed Sapiens Reference Sequence
SBP	Systolic blood pressure
SE	Standard error
SKAT	SNP-set (Sequence) Kernel Association Test (SKAT)
SKAT-O	Optimal Unified Test (SKAT-O)
SNV	Single nucleotide variants
TC	Total cholesterol
TOPMed	Trans-Omics for Precision Medicine
TRIG	Triglycerides
tRNA	Transfer ribonucleic acid
VAF	Variant allele fraction
WGS	Whole genome sequencing

CHAPTER 1 INTRODUCTION

1.1 Introduction

The human genome consists of nuclear DNA (nDNA) and mitochondrial DNA (mtDNA). In the last decades, thousands of genetic association studies have been conducted to identify the associations between nuclear DNA genetic variants and many disease traits.^{1;2} However, due to the limitation of conventional genotyping technology, the associations between disease traits and mitochondrial DNA genetic variations have not been explored exhaustively. Mitochondria are important organelles producing cellular energy in human cells. Through oxidative phosphorylation (OXPHOS), mitochondrion generates adenosine triphosphate (ATP), which is an organic compound that provides energy to drive many processes in living cells.³ Human mtDNA is a circular molecule of 16,569 base pairs, encoding 13 proteins in the energy production pathway, 22 transfer ribonucleic acid (tRNA), and two ribosomal RNA (rRNAs) for the biosynthesis of mitochondrion itself.⁴ Based on the similarity between mtDNA and the chromosome of proteobacteria, the endosymbiont hypothesis states that the mitochondrion evolved from a bacterial progenitor via symbiosis within an essentially eukaryotic host cell.⁵

Multiple mitochondria exist in a human cell and each mitochondrion contains multiple copies of mtDNA. Thus, unlike nuclear DNA with two copies per human cell, many copies of mtDNA molecules are present per human cell.⁶ The

number of copies of mtDNA varies widely by tissue and cell types in humans. For example, 4000-6000 copies of mtDNA molecules are present in tissues that need the most energy, such as cardiac and skeletal muscle. In contrast, a few hundred mtDNA molecules are present in blood cells.⁷ Although most mutations may only affect a small proportion of mtDNA copies, the increase in the number of mutant loci and a proportion of mtDNA copies with such mutations during aging may result in age-related diseases including cardiovascular disease and cancer.⁸

1.2 Nuclear Inheritance and Mitochondrial Inheritance

Human nDNA is packaged into thread-like structures called chromosomes. Normal human cells contain 23 pairs of chromosomes per cell. For each pair of chromosomes, one chromosome is inherited from the mother and another is inherited from the father through fertilization.⁹ On the contrary, mitochondria are inherited exclusively from the mother. The egg has thousands of mitochondria and they are passed to the developing embryo. In the sperm tail, mitochondria are used for propelling the sperm cells, and the tail is lost during fertilization.¹⁰ After fertilization, the mitochondria in sperm are usually destroyed by the egg cell.¹⁰ In addition, there is no clear evidence that recombination occurs in mtDNA.¹¹ Because the mtDNA genome is located near the oxidative phosphorylation system which produces reactive oxygen species such as superoxide and hydrogen peroxide,¹² mtDNA has a higher mutation rate than nuclear DNA. Thus, somatic mutations in mtDNA often occur and are

accumulated throughout the lifetime, and therefore, may result in age-related diseases.

1.3 Mitochondrial DNA Reference Sequences

The mtDNA variations are identified by comparing the sequence reads with a mtDNA reference sequence. Several mtDNA reference sequences have been commonly used. The first reference sequence is Cambridge Reference Sequence (CRS) which was assembled in 1981.¹³ This sequence was derived from sequencing the mitochondrial genome from one woman of European descent by Fred Sanger et al. at the University of Cambridge during the 1970s.¹³ However, eleven errors were found when resequencing the mtDNA from the same woman in 1992.^{14; 15} These eleven errors include one extra base pair in position 3107 and several incorrect assignments of single base pairs. The revised version of the sequence reference, designated as the revised Cambridge Reference Sequence (rCRS), was published by Andrews et al. in 1999.¹⁶ Haplogroups represent the major branch points on the mitochondrial phylogenetic tree.¹⁷ Because rCRS belongs to European haplogroup H2a2a1 which is not a maternal common ancestor of all living humans, some researchers argued that it may be better to choose a reference sequence starting with a “Mitochondrial Eve” (known as the maternal common ancestor of all living humans) as root for haplogroup prediction or comparing the changes in different haplogroups. Therefore, a new reference sequence, referred as Reconstructed

Sapiens Reference Sequence (RSRS), was assembled.¹⁸ The RSRS is a hypothetical sequence that uses both a global sampling of modern human samples and samples from ancient hominids.¹⁸ The RSRS represents the ancestral genome of Mitochondrial Eve. Nowadays debates remain about which reference sequence to choose between rCRS and RSRS. The favorer of RSRS claims that using rCRS may result in bias when tracing back from modern mtDNA sequences to our distant common maternal ancestor (the Mitochondrial Eve) because rCRS was from a European haplogroup.¹⁶ On the other hand, some researchers state that most of the potential rCRS-biased errors could have been prevented by referring to an mtDNA tree because the mtDNA classification tree could distinguish older inherited variants from recent or somatic ones which can help to avert potential oversights.¹⁹ Besides, additional reference sequences may confuse when adopting a phylogenetic approach.¹⁹ In addition to rCRS and RSRS, several alternative reference sequences have also been used in the field. These sequences include Yoruba, Uganda, Swedish and Japanese reference sequences.²⁰ Nevertheless, rCRS is the mostly used mtDNA reference in the field.

1.4 mtDNA Homoplasmy and Heteroplasmy

mtDNA variations are identified by comparing sequence reads with a reference sequence (e.g., rCRS or RSRS). Because multiple mtDNA copies are present in each cell, a mutant (alternative) allele may coexist with a wild-type

allele at the same mtDNA locus in the same cell or individual, and this phenomenon is called heteroplasmy.²¹ In contrast, a homoplasmic variant occurs when an alternative allele is presented in all copies of mtDNA in an individual. Several common homoplasmic variants have been identified, such as 8701G>A and 10398A>G.^{20; 22} Some of the homoplasmic variants are used to define mtDNA haplogroups. More than 4,000 different mtDNA haplogroups have been identified.^{17; 23} Some of the macro-haplogroups are commonly found in different populations. For instance, H, J, K, N1, T, U4, and U5 are common haplogroups in European populations; L0, L1, L2, L3, L4, L5, and L6 are common haplogroups in African populations; F, C, W, M, D, N, K, U are common haplogroups in Asian populations.¹⁷ Homoplasmy can be detected by conventional genotyping techniques. However, based on several previous studies²⁴⁻²⁶, most of the mtDNA heteroplasmic mutations are rare, only occurring in one or a few individuals. In addition, the mutant-to-wild type allele ratio in an individual is low for most heteroplasmic mutations. Therefore, mtDNA heteroplasmic mutations can only be optimally studied when deep sequencing of mtDNA is available in a large number of individuals.²⁵ Both mtDNA homoplasmic variants and heteroplasmic variants may be associated with disease traits.²⁷ More homoplasmic sites have been found in the human population than heteroplasmic sites.²⁵

1.5 Association Analysis of mtDNA Variants

Due to the low coverage of conventional genotyping technologies, only homoplasmic variants and high-level heteroplasmic variants can be detected accurately around one decade ago.²⁵ Several common homoplasmic variants (e.g. 15326A>G, 8860A>G) and heteroplasmic mutations (e.g. 16093T>C, 16129G>C) were identified.²⁰ The common homoplasmic variants are well studied. The chi-squared test is commonly used for association testing of common homoplasmic variants with disease traits in case-control studies.²⁸ Linear models and linear mixed models (LMM) are widely used to analyze the associations between common homoplasmic variants and continuous traits in unrelated samples and correlated samples, respectively.²⁹⁻³² Logistic regression models and generalized linear mixed models (GLMM) are commonly employed for analyzing binary traits in association analyses with common homoplasmic variants.³³ However, most heteroplasmic mutations are rare in the sense that, for most heteroplasmic mutations, the heteroplasmy level is low in an individual, and most heteroplasmic mutations only occur in one or a few individuals in the human population. Therefore, appropriate statistical methods to investigate relationships between rare heteroplasmic mutations and disease traits remain to be studied.

1.6 Pipelines for the Identification of mtDNA Sequencing Variations

Conventional Sanger sequencing lacks sensitivity for detecting mtDNA mutations with heteroplasmy levels lower than 20%.³⁴ Thanks to deep sequencing techniques, low-level heteroplasmic variants can be identified precisely. The limit of detection of pyrosequencing methods is close to 5%^{35; 36} (i.e. only mutations with heteroplasmy level $\geq 5\%$ can be detected), and close to 1% by reversible terminated chemistry^{37; 38}. The duplex sequencing technology improves the limit of detection further to 0.01%.^{39; 40} With a greatly reduced cost in next-generation sequencing, mtDNA sequences have become available in a large human population through the whole genome sequencing (WGS) effort by the Trans-Omics for Precision Medicine (TOPMed) project supported by National Heart, Lung, and Blood Institute (NHLBI).⁴¹ The average coverage of nuclear DNA is > 30 fold in WGS and the corresponding average coverage of mtDNA by WGS is around 1000-4000 fold. Several bioinformatics pipelines have been developed to identify mtDNA sequence variations. Several widely used pipelines include MToolBox⁴², mitoCaller²⁴, Mutect2⁴³, and Mitomaster⁴⁴. MToolBox is a commonly used software for mtDNA variation identification, annotation, and haplogroup prediction. mitoCaller is also a computationally efficient pipeline to detect mtDNA variants by using a likelihood-based model, while the algorithm of Mutect2 is based on a pair-hidden Markov model probabilistic model for variation recognition. In chapter 3, we describe the implementation of MToolBox in the identification of heteroplasmic variants.

1.7 Gene-based Association Tests for Nuclear DNA

Genome-wide association studies (GWAS) have identified thousands of associations between disease traits and common variants of nuclear DNA (nDNA) with minor allele frequencies (MAFs) $\geq 1\%$.^{1; 2} Previous studies showed that low frequency and rare variants (MAFs $< 1\%$) contribute a considerable proportion of the heritability of phenotypes in addition to common variants, where heritability is a measure that estimates the degree of variation in a phenotype between individuals that is due to genetic variation in the population.⁴⁵ However, the single-variant association tests are, in general, underpowered to detect associations of phenotypes and rare variants with MAFs $< 1\%$ in conventional cohort studies.

Several variant-set tests (known as aggregate unit tests) have been developed to increase the power in association testing of rare variants by aggregating multiple rare variants in a region or a gene. The burden test⁴⁶ and sequence kernel association test (SKAT)⁴⁷ are two widely used variant-set tests in association analyses. Both methods are suitable for continuous, binary, and time-to-event outcomes and incorporate covariates. Burden tests are more powerful if a large proportion of variants in the region are associated with phenotypes, and all the allele effects are in the same direction.⁴⁷ But the power is markedly reduced if many genetic effects have the opposite direction. In contrast, SKAT outperforms the burden test if a small proportion of variants is causal

and/or the effect directions differ.⁴⁷ The burden test and SKAT can be combined by calculating a weighted average of their test statistics as $Q_\rho = \rho Q_{burden} + (1 - \rho)Q_{SKAT}$, which follows a mixture of chi-square distributions asymptotically. SKAT-O⁴⁸ is constructed to identify the optimal weight (combination) to minimize the p-value over different combinations. SKAT-O is adaptive for different scenarios of causal variants in the tested region. For instance, SKAT-O provides robust statistics irrespective of the proportion of causal variants within a region and the directionality of the causal effects. In practice, we rarely have information about the optimal combination. Therefore, a grid search over ρ is needed to find the combination with a minimal p-value for the statistic Q_ρ . The Q_ρ statistic controls for multiple testing by a one-dimensional numeric integration. Therefore, SKAT-O is more computationally intensive than the burden test and SKAT.⁴⁸

Different weights can be assigned to the genetic variants in a region. The weight of each variant may be based on the MAF or the functional score such as the Combined Annotation Dependent Depletion (CADD) score. For example, one can assign larger weights to the variants with larger CADD scores when assuming that larger functional scores lead to larger effects on disease traits. An appropriate choice of weights for causal variants may increase power in association analyses. In SKAT, beta density functions are widely used as a group of flexible weight functions that depends on MAF, such as $w_j = \text{Beta}(MAF_j, a_1, a_2)$, where a_1 and a_2 are pre-specified parameters, and MAF_j is

the MAF of the j -th variant.⁴⁷ The parameters are set to be $a_1 = 1$ and $a_2 = 1$ if we assume that all the variants have the same weights. Another common choice is $a_1 = 1$ and $a_2 = 25$ by assuming that more rare variants have larger effects on the trait.⁴⁷ However, the optimal weights are unknown in practice.

An omnibus test, the aggregated Cauchy association test (ACAT)⁴⁹, is used to combine different tests (e.g., burden and SKAT) with different weights. Because the distribution of the test statistics of ACAT is well approximated by a Cauchy distribution under the null hypothesis, the correlation structure of individual variant-set tests statistics is not needed to calculate the p-value.⁴⁹ Hence the ACAT test is fast to compute once we have the p-values of the individual variant-set tests we wish to combine.

1.8 Outline of Dissertation

In the last decades, thousands of genetic association studies have been conducted to identify the associations between genetic variants in nuclear DNA and many disease traits. However, due to the limitation of conventional genotyping technology, the associations between disease traits and mitochondrial DNA genetic variations have not been explored exhaustively. To this end, we develop a novel framework for the association analysis of rare (MAF < 0.01) heteroplasmic mitochondrial DNA mutations with disease traits.

In chapter 2, we establish a statistical framework that incorporates a pre-specified threshold for identifying heteroplasmic variants and performs association analyses with a few methods, including the original burden test⁴⁶ with its extensions including adaptive burden test, Z-score weighting approach, and variable threshold approach,^{50; 51} and SKAT.⁴⁷ In the framework, we use two definitions to define heteroplasmic variants. This framework also uses an aggregated Cauchy association test (ACAT-O)⁴⁹ and SKAT-optimal (SKAT-O)⁴⁸ to combine information from multiple gene-based methods applied to assess the association of phenotypes with heteroplasmic variants. Furthermore, this framework can easily incorporate different types of weights (e.g., a function of the variant allele fraction and the predicted functional score). We evaluate the performance of these methods with continuous and binary traits with a comprehensive simulation study. We also apply the methods to real data to assess the association of heteroplasmy with age and sex in several large cohorts with whole genome sequencing data.

In chapter 3 we propose a clustering method for trait-associated rare variant identification and effect direction classification based on a Gaussian mixture model (GMM). Without loss of generalizability, this proposed method applies to both nuclear and mitochondrial rare variants. First, we perform ACAT-O combining burden test and SKAT to detect regions in which rare variants show associations with a trait. For a given region associated with a disease trait, we fit

single variant models to obtain association statistics between phenotype and rare variants within the region. Based on the effect size, standard error (SE), and effect direction from the single variant model, this novel clustering method may identify risk and/or protective rare genetic variants with a trait of interest. The single variant model will be described in detail in Chapter 3. Because the performance of GMM is sensitive to the initialization of parameters, we calculate the initial proportion values for associated variants by an optimal combination of z-scores. In addition, the method may also cluster the associated rare variants into potential subgroups based on the direction and magnitude of effect sizes of those rare variants in association analyses. We evaluate the performance of the proposed clustering method with a simulation study and apply it to a real dataset.

In chapter 4, we apply the framework we develop in Project 1 to explore the associations between heteroplasmic variants and several cardiometabolic disease (CMD) traits for five large cohorts of ARIC, CHS, FHS, JHS, and MESA. The CMDs include obesity, diabetes, hypertension, and hyperlipidemia as well as several continuous traits defining these CMD diseases. We incorporate three coding definitions of heteroplasmy including the two definitions defined in project 1, and an additional definition based on Mito-Score that incorporates functional annotation and region constraint to evaluate a heteroplasmic variant. We perform cohort- and ancestry-specific association analyses. We meta-analyze the results across cohorts using a fixed-effect model (for the burden test) and Fisher's

method (for the other tests). We summaries our results, make conclusions and outline future work in chapter 5.

CHAPTER 2 ASSOCIATION ANALYSIS OF MITOCHONDRIAL DNA HETEROPLASMIC VARIANTS: METHODS AND APPLICATIONS

2.1 Introduction

Mitochondria are important organelles producing cellular energy through oxidative phosphorylation, calcium homeostasis, regulation of innate immunity, programmed cell death, and stem cell regulation.⁴ The maternally inherited mitochondrial genome is a circular molecule of double-stranded DNA (mtDNA). Human mtDNA consists of 16,569 base pairs and is essential for proper mitochondrial function. mtDNA encodes 22 tRNAs and 2 rRNAs, and 13 proteins that are involved in the energy production pathway.⁴ Hundreds to thousands of mtDNA molecules are present per human cell, depending on the cell's energy requirement.⁶ Heteroplasmy refers to a phenomenon where two or more alleles coexist at the same site in a mixture of mtDNA molecules within a cell or an individual.²¹ Based on our previous studies⁵² and other studies,²⁴⁻²⁶ 98% of mtDNA heteroplasmic variants are rare, only present in one (i.e., singleton) or a few individuals. In addition, most heteroplasmic variants display low VAFs in the general human population.^{24-26; 52} Nonetheless, the increase in both their number and VAFs of heteroplasmy during aging may contribute to age-related diseases, including cardiovascular disease and cancer.^{6; 8}

With a greatly reduced cost in next-generation sequencing technologies, hundreds of thousands of human genome samples, including mtDNA, have been

sequenced. The availability of mtDNA sequences with high coverage (e.g., > 2000-fold) in large human populations⁵² provides for the detection of rare, low-level heteroplasmic variants that are potentially associated with disease traits. The commonly used statistical methods to analyze rare variants in the nuclear genome, e.g., burden tests⁴⁶ and the sequence kernel association test (SKAT),⁴⁷ have not been evaluated for their performance in the context of ultra-rare variants such as mitochondrial heteroplasmic variants. In addition, there is no standard procedure or approach for the analysis of heteroplasmic variants. Therefore, it is important to develop a novel framework for testing the association of heteroplasmic variants with disease traits.

We propose a statistical framework for association analysis of heteroplasmic variants with a trait. This framework incorporates a pre-specified threshold for identifying true heteroplasmic variants and performs association analyses with a few methods, including the original burden test⁴⁶ and its extensions,^{50; 51} and SKAT.⁴⁷ This framework also uses an aggregated Cauchy association test (ACAT-O)⁴⁹ and SKAT-optimal (SKAT-O)⁴⁸ to combine information from multiple gene-based methods applied in association analyses. Furthermore, this framework can easily incorporate different types of weights (e.g., the variant allele fraction and the predicted functional score). In this study, we evaluate the performance of these methods using simulated and real data to assess the association of heteroplasmy with continuous and binary traits.

2.2 Methods

2.2.1 Definition of mtDNA Sequence Variants

Variant alleles are identified by comparing sequence reads in mtDNA to reference sequence, e.g., the revised Cambridge Reference Sequence (rCRS)¹⁶ or Reconstructed Sapiens Reference Sequence (RSRS)¹⁸. A variant allele fraction (VAF) is the proportion of the variant alleles over all sequence reads observed at an mtDNA site in an individual. To minimize false positive findings, a heteroplasmy is defined by a pre-specified threshold $\tau = (\tau_1, \tau_2)$. Let VAF_{ij} be the VAF of a variant at mtDNA site j^{th} in the i^{th} individual. Here $j = 1, \dots, m$, and $i = 1, \dots, n$. A site j is not considered as a variant if $VAF_{ij} < \tau_1$ in individual i ; it is considered as a heteroplasmy if $\tau_1 \leq VAF_{ij} \leq \tau_2$; and it is considered as a homoplasmy if $VAF_{ij} > \tau_2$.

Let G_{ijt} be the coding of the heteroplasmy at the j^{th} site of i^{th} individual with a VAF threshold τ . We consider two coding schemes for variants in association testing. First, we define a heteroplasmic variant by an indicator function in which the VAF of a heteroplasmy is not incorporated:

$$G_{ij\tau} = 1_{\tau}(VAF_{ij}) = \begin{cases} 1 & \text{if } VAF_{ij} \in \tau \\ 0 & \text{o.w.} \end{cases} \quad (\text{Definition 1})$$

Second, we define a heteroplasmy by incorporating its VAF:

$$G_{ij\tau} = \begin{cases} VAF_{ij} & \text{if } VAF_{ij} \in \tau \\ 0 & \text{o.w.} \end{cases} \quad (\text{Definition 2})$$

Because definition 2 results in distinct scales between heteroplasmic sites, we standardize the coding of each heteroplasmy.

2.2.2 A Phenotype Model

For subject i , let y_i denote a phenotype with mean μ_i . Let $X_i = (X_{i1}, \dots, X_{iq})^T$ denote a vector of covariates, and let $G_{i\tau} = (G_{i1\tau}, \dots, G_{im\tau})^T$ be a vector of the coding of m mtDNA heteroplasmic variants in a region or gene. We consider a generalized linear model (GLM) framework to investigate the relationship between a set of mtDNA heteroplasmic variants in a region or gene and a phenotype.⁵³

$$g(\mu_i) = d_0 + X_i^T \mathbf{d} + M_i + N_i + G_{i\tau}^T \boldsymbol{\beta} \quad (\text{Equation 1})$$

where $g(\mu_i) = \mu_i$ for a continuous trait and $g(\mu_i) = \text{logit}(\mu_i)$ for a binary trait. To be more generalizable, we let M_i be a polygenic component of the mtDNA, and N_i be a polygenic component of the nuclear genome.³¹ For family data, M_i and N_i represent random effects from maternal and nuclear correlation matrix, respectively. In Equation 1, d_0 is an intercept, $\mathbf{d} = (d_1, \dots, d_q)^T$ is a column vector of the effects from covariates, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_m)^T$ is a column vector of the effects from a set of heteroplasmic variants. The testing of the null hypothesis of no association between mtDNA sequence variations and a trait is equivalent to testing $H_0: \boldsymbol{\beta} = (\beta_1, \dots, \beta_m) = \mathbf{0}$. The score statistic for mutation j is defined as

$$U_j = \sum_{i=1}^n G_{ij\tau} (y_i - \hat{\mu}_i) \quad (\text{Equation 2})$$

where $\hat{\mu}_i$ is the estimated mean of y_i under the null hypothesis ($H_0: \beta = 0$) by fitting the null model $g(\mu_i) = d_0 + X_i^T \mathbf{d}$.

2.2.3 Gene-based Tests

The original burden and SKAT

In this study, we only considered heteroplasmic variants with population-level frequency $MAF_{Hj} = \frac{1}{n} \sum_i^n G_{ij\tau} < 0.01$ in association analyses. Here, n is the number of participants in the study and MAF_{Hj} refers to the minor allele frequency (MAF) of a heteroplasmic variant (H) j . The original burden test⁴⁶ (referred to as Burden) and SKAT⁴⁷ are often used to aggregate the effects of rare variants in a genetic region in autosome. The corresponding test statistic for Burden is $Q_{burden} = (\sum_{j=1}^m w_j U_j)^2$. Under the null, Q_{burden} follows a chi-square distribution asymptotically with 1 degree of freedom. The SKAT method⁴⁷ uses a variance component framework and the corresponding test statistic is $Q_{SKAT} = \sum_{j=1}^m w_j^2 U_j^2$. This test statistic follows a mixture of independent chi-square distributions asymptotically with 1 degree of freedom under the null. For both Burden and SKAT statistics, w_j is a weight⁵³ that an investigator may choose for mutation j .

Extensions to the original burden test

The burden test has larger power than SKAT when rare variants in a gene/region display the same effect direction on a trait⁴⁸. Therefore, the adaptive burden test⁵⁰ (denoted as Burden-A) was proposed to improve the power for Burden by incorporating the effect directions of rare variants within the target gene region. The adaptive burden test⁵⁰ (denoted as Burden-A) is an extension of the Burden test by changing the coding sign of single nuclear variants (SNVs) based on an arbitrary threshold p-value in regression analysis and the observed direction of effect. This test may have an advantage over the Burden⁴⁶ test by selecting possible causal variants to be included in the test. The steps that perform the adaptive burden test⁵⁰ for association analyses of heteroplasmic variants are described as the following.

Step 1: For each heteroplasmic variant j , we fit a single mutation model described below

$$g(\mu_i) = \alpha_0 + Age_i\beta_1 + Sex_i\beta_2 + G_{i,j}\beta_{3,j} \text{ (Equation 3)}$$

where $g(\cdot)$ is the link function connecting the μ_i which is the conditional mean of phenotype y_i , to age, sex, and genetic dosage ($G_{i,j}$) of j^{th} heteroplasmy of the i^{th} subject. Here α_0 is the intercept, β_1 and β_2 are the beta coefficients of age and sex, $\beta_{3,j}$ is the genetic effect of the j^{th} heteroplasmy in the single mutation model.

We obtain the estimate of the genetic effect $\widehat{\beta}_{3,j}$ and the p-value $p_{M,j}$ for $j=1 \dots J$.

Step 2: Given a pre-specified p-value cutoff p_c , we change the coding sign of the heteroplasmy j if $p_{3,j} \leq p_c$ and $\widehat{\beta}_{3,j} < 0$. The signs of the other heteroplasmic variants remain the same. Hence, we obtain a new genetic dosage matrix G^{new} with the same dimensions as the original genetic dosage matrix. We set $p_c = 0.1$ in our study based on a previous study that proposed the adaptive burden method (Burden-A).⁵⁰

Step 3: Perform the original burden test with the new genetic dosage matrix G^{new} and obtain the p-value, p_{new} .

Step 4: Permute the phenotype $\{Y_i\}$ B times to obtain B sets of permuted data $\{(Y_i^{(b)}, X_i, G_i)\}$ for $b=1, \dots, B$. For each permuted data, we repeat steps 1-3 and obtain a p-value $p^{(b)}$. Therefore, we generate an empirical null distribution of $p_{new}: \{p^{(b)}\}$ with $b=1, \dots, B$. The empirical p-value of the test is calculated as $\sum_{b=1}^B I(p^{(b)} < p^{new}) / B$. We choose $B=50000$ for α level of 0.001.

Burden-A⁵⁰ uses the same weight for all the variants, which may lead to power loss. Sha and Zhang⁵¹ proposed a z-score weighting approach (referred to as Burden-S) to minimize this limitation. According to the method by Sha and Zhang⁵¹, let z_j denote the z-score of j^{th} heteroplasmic variant from Equation 3,

where $z_j = \frac{\widehat{\beta}_{3,j}}{SE(\widehat{\beta}_{3,j})}$. The weight of the j^{th} heteroplasmy w_j is set to be z_j . The weight matrix is formed as

$$W = \text{diag}(w_1, \dots, w_J)$$

and the score-weighted genetic dosage matrix is defined as

$$G^S = GW$$

In analogy to the adaptive burden test, an empirical p-value is evaluated by a permutation test for the z-score weighting method. Note that we assign larger weights to heteroplasmic variants with larger z-scores. The sign of the heteroplasmic variants with negative z-scores is switched. Because switching the sign of all heteroplasmic variants with negative z-scores would result in extreme p-values under the null hypothesis, the null distribution would have a heavy tail and may lead to power loss. Some heteroplasmic variants may give rise to extreme z-scores by chance and this may result in an inflated type I error rate.

To avoid such a situation, we modify the z-score weights to have lower (z=-1.5) and upper (z=1.5) bounds. That is, we set $w_j = 1$ if $|z_j| < Z_{0.05}$ where $Z_{0.05} \approx 1.65$, the 95 percentile of the standard normal distribution. If $z_j \geq Z_{0.05}$, w_j is assigned to be $z_j - Z_{0.05} + 1$, with an upper limit of 1.5. Similarly, if $z_j \leq -Z_{0.05}$, w_j is assigned to be $z_j + Z_{0.05} - 1$, with a lower limit of -1.5.

Another limitation of Burden-A is that the cutoff p_c based on marginal models is chosen arbitrarily. To overcome this, Sha and Zhang proposed the variable threshold approach⁵¹ (Burden-V) that searches for an optimal cutoff in Burden-A based on Equation 3. This method is implemented in the following steps:

Step 1: We select various percentiles q^S of the p-values $S = \{p_{3,j}, j = 1 \dots J\}$ from Equation 3 as the thresholds. According to Sha and Zhang⁵¹, they choose all possible p-values as the candidate thresholds, which leads to an intensive computational burden. Therefore, we choose the 15th, 30th, 50th, 70th, and 85th percentiles to be the thresholds with a continuous trait, denoted by q_{15}^S , q_{30}^S , q_{50}^S , q_{70}^S , q_{85}^S .

Step 2: For a given percentile q^S , we only include heteroplasmic variants with the p-value $p_{3,j} \leq q^S$ and change the sign of the heteroplasmic variant in G if the corresponding beta coefficient $\widehat{\beta}_{3,j} < 0$. Thus, we obtain a manipulated genetic dosage matrix G_{q^S} .

Step 3: We perform the original burden test based on G_{q^S} and get the p-value. The five percentiles yield five p-values: $p_{q_{15}^S}$, $p_{q_{30}^S}$, $p_{q_{50}^S}$, $p_{q_{70}^S}$ and $p_{q_{85}^S}$. We

also run a burden test by the original genetic dosage matrix G and get the p-value p_0 . Based on these six p-values $K = \{p_0, p_{q_{15}^S}, p_{q_{30}^S}, p_{q_{50}^S}, p_{q_{70}^S}, p_{q_{85}^S}\}$, we define two test statistics that are referred to as the Burden-V1 and Burden-V2 methods.

$$T_1 = \min K \text{ (Burden-V1)}$$

$$T_2 = \sum_{p \in K} \tan((0.5 - p)\pi) / |K| \text{ (Burden-V2)}$$

where T_2 is the test statistic of ACAT⁴⁹.

Because most of the heteroplasmic variants are singletons, the regression model, Equation 3, of logistic regression with a binary trait leads to biased estimates and an extremely conservative p-value $p_{3,j}$ (>80% of the p-values > 0.9). Hence, we modify the single mutation model, and fit a logistic regression under the null hypothesis of no genetic effect on the trait and obtain the residuals:

$$\text{logit}(\mu_i) = \alpha_0 + \text{Age}_i \beta_1 + \text{Sex}_i \beta_2$$

These residuals are rank-base inverse normalized. Then we regress the transformed residuals on each of the heteroplasmic variants to get the p-value and beta coefficient. In addition, because a logistic regression with few rare mutations may lead to conservative results, we set the thresholds to be 50th, 70th, and 85th percentiles.

The weights

The weights, $\beta(MAF, 1, 25)$, are used to put more weight on rarer variants in association analyses.⁴⁷ However, 98% of mtDNA heteroplasmic variants are only present in one (i.e., singleton) or a few individuals.^{24-26; 52} Due to this extreme rareness, the $\beta(MAF, 1, 25)$ weights play a minimum role in the association analysis of rare heteroplasmic variants. For example, applying the $\beta(MAF, 1, 25)$ weights to analyzing a singleton heteroplasmy or a heteroplasmic variant of five individuals gives rise to almost identical weights (24.8 versus 24.0, respectively) in a cohort of 3,000 individuals. Simulation studies confirm the theoretical calculations (Supplementary Figure 2.6). Therefore, the $\beta(MAF, 1, 25)$ weights are not used in evaluating type I error and power.

2.2.4 Combining Multiple Gene-based Tests in a Gene/Region

We adopt two omnibus tests (ACAT-O and SKAT-O) to combine information from association testing of heteroplasmic variants with Burden and SKAT. A generalized SKAT (SKAT-O)⁴⁸ is constructed as a linear combination of an original burden test and a SKAT⁴⁸. The test statistic is $Q_\rho = \rho Q_{burden} + (1 - \rho)Q_{SKAT}$, where $\rho \in [0,1]$ is a weight applied to the original burden test statistic. Under the null hypothesis, Q_ρ follows asymptotically a mixture of independent chi-square distributions with 1 degree of freedom. The SKAT-O can be

constructed if we choose the minimum p-value of the different choices of ρ . The test statistic of SKAT-O is $Q_{SKAT-O} = \min \{p_{\rho_1}, \dots, p_{\rho_k}\}$. The significance of Q_{SKAT-O} can be assessed by a one-dimensional numerical integration. In our study, to make it comparable to ACAT-O, we take $\rho=0$ and 1 to combine Burden and SKAT. That is, $0 = \rho_1 < \rho_2 = 1$.

In ACAT-O⁴⁹, the test statistic of a region/gene is defined as $Q_{omnibus} = \frac{1}{2} [\tan\{(0.5 - p_{burden})\} + \tan\{(0.5 - p_{SKAT})\}]$. Here p_{burden} and p_{SKAT} denote the p-values from Burden test and the SKAT. Because this test statistic approximately follows a standard Cauchy distribution⁴⁹, the p-value of the test statistic can be approximated by $p_{omnibus} \approx \frac{1}{2} - \frac{\arctan(Q_{omnibus})}{\pi}$. The ACAT-O method is computationally fast and it efficiently combines p-values from individual tests of different methods when multiple weighting schemes are applied. Based on previous studies, we use $\alpha=0.001$ to control for multiple testing in association analyses across multiple genes/regions.³⁰

2.2.5 A Simulation Study

According to Equation 1, we simulated a continuous trait and a binary trait in response to heteroplasmic sites located in the mitochondrial Cytochrome b (MT-CYB) gene in European American participants (N=3,415) of the Atherosclerosis Risk in Communities (ARIC) Study⁵⁴. We only consider

independent samples and no correlation between nDNA and mtDNA, therefore, Equation 1 is simplified to $g(\mu_i) = d_0 + X_i^T \mathbf{d} + G_{i\tau}^T \boldsymbol{\beta}$ by setting M_i and N_i to 0. The CYB gene has a length of 1141 base pairs, and it is the fourth longest gene among the 13-mtDNA coding genes. Heteroplasmy was identified by a widely-used software package, MToolBox,⁴² with WGS data from TOPMed Freeze 8, released in February 2019, GRCH38.⁴¹ The rCRS¹⁶ was used to identify heteroplasmic variants. The CYB gene contains 121 heteroplasmic sites in European American participants in ARIC. Of those, 66 are nonsynonymous and rare variants. The traits were simulated as a function of these 66 nonsynonymous mutations (Supplementary Table 2.1) according to Equation 1. We simulated 50,000 replicates to evaluate the performance of the proposed methods with empirical type I error rate and power at $\alpha=0.001$. Because the simulated phenotype follows the same distribution across the 50,000 simulation replicates under null, we simulated one empirical null distribution using 50,000 permutations for each method with each coding definition.

Type I error rate

To evaluate the type I error rate, we simulated a phenotype, y , by Equation 1 under the null hypothesis: $y = 0.08Age + Sex + \varepsilon$, where $\varepsilon \sim N(0, 0.7)$. We applied a cutoff of 80% quantile to the simulated continuous phenotype to obtain a binary phenotype with a 20% prevalence rate. The observed type I error

rate is defined as the proportion of simulation replicates with p-values ≤ 0.001 under the null. We evaluated the type I error rate with a ratio of the observed type I error rate divided by 0.001. We calculate the 95% confidence interval of the empirical type I error rate of each method by the exact method using “binom.confint” function of the binom package in R. We define the empirical type I error rate to be conservative for a test if the upper limit of its 95% CI is smaller than 1. Analogously, we define the empirical type I error rate to be inflated for a test if the lower limit of its 95% CI is larger than 1. The empirical type I error rate is controlled well if its 95% CI contains 1.

Power estimation

To evaluate power, we simulated a continuous phenotype using the following model, a special case of Equation 1, with a genetic effect from sequence variations in the *CYB* gene: $y = 0.08Age + Sex + \mathbf{G}_\varphi^{cT} \boldsymbol{\beta} + \varepsilon$, where $\mathbf{G}_\varphi^{cT} = (G_{1\varphi}^c, \dots, G_{k\varphi}^c)$ is a vector that includes the coding for k randomly chosen causal heteroplasmic variants in the *CYB* gene. If we use a coding definition in simulating a phenotype, we use the same coding definition in data analysis. That is, we use the same coding definition in the phenotype simulation step and the analysis step. $\boldsymbol{\beta}^T = (\beta_1, \dots, \beta_k)$ is a vector of fixed effects for the selected causal mutations. We also applied a cutoff of 80% quantile to the simulated continuous phenotype to obtain a binary phenotype. The effect size of heteroplasmic variant

j is specified by $|\beta_j| = \sqrt{\frac{c}{\text{var}(G_{j\tau})}}$, where c is a constant defined as $c = \frac{R^2}{V^T D V}$. Here $\text{var}(G_{j\tau})$ is the variance of the heteroplasmic variant j , R^2 is the proportion of variance explained by all causal mutations, D is the correlation matrix between mutations, and V is a vector of the signs of β . The proportion of variance (R^2) explained by the causal heteroplasmic variants was set to be 1% for the continuous phenotype, and 2% for the binary phenotype. The variance explained by age, sex, and random error is around 24%, 25%, and 50% respectively. We calculate empirical power by adjusting for the empirical type I error rate. We also calculate empirical power not adjusting for empirical type I error rate.

Scenarios

In practice, it is possible that a proportion of heteroplasmic variants in an mtDNA gene have effects on a phenotype while the rest of heteroplasmic variants in the same gene have no effects (i.e., $\beta = \mathbf{0}$) or display opposite effects (i.e., some elements of β are positive and some others are negative) on the phenotype. We considered 5%, 25%, 50%, and 80% of the nonsynonymous heteroplasmic variants in the *CYB* gene to be causal and 50%, 80%, and 100% of these causal heteroplasmic variants to have the same effect direction with a phenotype. Therefore, we evaluated the power under 12 scenarios that vary the proportion of heteroplasmic variants to be causal and vary the proportion of the causal mutations to have the same directionality. To account for an inflated or

conservative type I error rate, we estimate empirical power as the proportion of p-values that is smaller than 0.1 quantile from all simulation replicates. We also present the power based on the nominal α level of 0.001. That is the proportion of simulation replicates with $p\text{-values} \leq 0.001$.

2.2.6 Application of the Proposed Framework in Real Data

Study participants

We applied the framework to analyze heteroplasmy with traits in five large cohorts, including ARIC⁵⁴, Framingham Heart Study (FHS)⁵⁵⁻⁵⁷, Cardiovascular Health Study (CHS)⁵⁸, Jackson Heart Study (JHS)⁵⁹ and Multi-Ethnic Study of Atherosclerosis (MESA) (Table 2.2).⁶⁰ These cohorts are prospective cohort studies designed to investigate cardiovascular disease and its risk factors across different US populations. Participants in these five cohorts received whole genome sequencing (WGS) with an average coverage of 39-fold from the Trans-Omics for Precision Medicine (TOPMed) program, sponsored by the National Institutes of Health (NIH) National Heart, Lung and Blood Institute (NHLBI).⁴¹ We excluded two duplicated individuals from JHS and eight from MESA because they overlapped with ARIC.

Identification of mtDNA heteroplasmy

Quality control of WGS sequencing was described previously⁵². We applied MToolBox⁴² to all participating cohorts (WGS TOPMed Freeze 8, released in February 2019, GRCH38)⁴¹ with rCRS.¹⁶ We applied the 3%-97% of thresholds to identify heteroplasmy. The selection of the 3%-97% threshold with TOPMed WGS data and the detailed information for quality control of mtDNA sequence variations was described previously.⁵²

Association analysis of heteroplasmy

We aim to identify age-associated mtDNA gene heteroplasmic variants and explore the association between sex and heteroplasmy. We applied two coding definitions to evaluate these gene-based tests and omnibus methods to analyze heteroplasmic variants in cross-sectional association analyses of age (a continuous trait) and sex (a binary trait) with rare heteroplasmy ($MAF_{Hj} < 0.01$) in 16 genes/regions using the GLMM⁶¹ framework with heteroplasmic variants in sixteen genes/regions as the predictor variables. For illustration purposes, we use age and sex as outcome variables. We performed cohort-specific and ancestry-specific association analyses with gene-based tests and omnibus tests. A previous study showed that the year of blood draw was significantly associated with mtDNA copy number.⁶² We tested if the heteroplasmic burden was associated with a year of blood draw in each cohort (Supplementary Table 2.3). We adjusted for the year of blood draw in cohort-specific association analyses if

the year of blood draw showed a significant association with heteroplasmic burden in a cohort. Whole genome sequencing was conducted using whole blood-derived DNA. A previous study showed that white blood cell count and platelet count were associated with the total heteroplasmic burden,⁵² therefore, we performed sensitivity analyses with additional adjustments for white blood cell count and differential count (the proportions of neutrophil, lymphocyte, monocyte, eosinophil, and basophil) and platelet variables. Because cell count/platelet variables were available only in a subset of cohort/participants, the sensitivity analyses were performed in FHS (n=2551) and JHS (n=2737). We compared the regression estimates between the models with and without cell count/platelet variables in the 2551 FHS participants and also in the 2737 JHS participants.

Ancestry-specific meta-analysis was performed separately in participants of European ancestry and African ancestry. Meta-analysis was also performed to combine the results from each ancestry. We used two methods to conduct meta-analyses. We first combined the p-values across cohorts by Fisher's method which does not employ weight in meta-analyses.⁶³ In addition, we performed a meta-analysis using the fixed-effects inverse variance method⁶⁴ to combine the summary statistics of the original burden tests. Here, we hypothesized that there is only one true treatment effect for the association of heteroplasmy with a trait between studies. We presented a meta-analysis of all participants as the main result. We used α level of 0.001 for significance³⁰ in association testing. We

performed 50,000 permutations to obtain empirical p-values for all of the burden-extension methods. All analyses in simulation and application used R software version 3.6.0.⁶⁵

2.3 Results

We simulated a continuous trait and a binary trait based on heteroplasmic sites located in the mitochondrial Cytochrome b (MT-CYB) gene which is the fourth longest gene of mtDNA (Methods & Supplementary Table 2.1). Below we presented results from simulation studies to evaluate type I error rate and power for several gene-based tests and the two omnibus tests. Of note, we used the same coding definition in simulating a phenotype and in estimating type I error rate and power. We also presented the findings from the application of these methods to real data in the five large cohorts with WGS.

2.3.1 Empirical Type I Error Rate of Simulation Studies

We employed two coding definitions of heteroplasmy which were described thoroughly in the method section. By definition 1 with a continuous trait, the type I error rate was appropriately controlled for Burden, Burden-A, Burden-S, Burden-V1, and ACAT-O because the 95% CI of their empirical type I error rate contains the null value of 1. The empirical type I error rate of burden-V2 was inflated because the lower limit of its 95% CI (1.23) was larger than the null

value of 1. The SKAT and SKAT-O displayed a conservative empirical type I error rate because the upper limit of their 95% CI (0.9) was smaller than the null value of 1 (Table 2.1). For a binary trait with a prevalence of 20%, the type I error rate was well controlled for Burden, Burden-S, SKAT, SKAT-O, and ACAT-O because the 95% CI of their empirical type I error rate included 1 (Table 2.1).

By definition², the type I error rate was properly controlled for Burden, Burden-A, Burden-V1, SKAT-O, while ACAT-O. Burden-S and Burden-V2 had an inflated type I error rate. The type I error rate of SKAT was conservative. (Supplementary Table 2.2). With a binary trait, Burden, Burden-A, Burden-S, and ACAT-O controlled the type I error rate well. The empirical type I error rate of Burden-V1 and Burden-V2 was extremely conservative. The SKAT and SKAT-O had moderately conservative type I error rates (Supplementary Table 2.2).

Of note, the weight, $\beta(\text{MAF}, 1, 25)$, which is widely used in gene-based association testing of rare variants in nDNA showed minimum effects to upper weight the rarer heteroplasmic variants in association testing. That is, the $\beta(\text{MAF}, 1, 25)$ and $\beta(\text{MAF}, 1, 1)$ provided similar results. Therefore, this weight was not evaluated in the subsequent results. We used equal weights for each heteroplasmic variant at the population level for all our analyses.

2.3.2 Empirical Statistical Power of Simulation Studies

Gene-based tests by Definition 1

We estimated empirical power using the proportion of p-values that is smaller than 0.1 quantile in simulation studies (Figure 2.1, Supplementary Figure 2.1) and using the fixed (i.e., nominal) $\alpha = 0.001$ (Supplementary Figure 2.2-2.3). For both continuous and binary traits, as expected, for all gene-based tests, power was improved when the proportion of causal variants (of all variants) increased and/or when the proportion of causal variants with the same effect direction increased for both definitions (Figure 2.1, Supplementary Figure 2.1-2.3).

When 100% of heteroplasmic variants had the same effect direction, Burden and the burden extension methods displayed comparable power, adjusting for empirical alpha rate (Figure 2.1, Supplementary Figure 2.1). However, when any proportion of the causal variants displayed different effect directions, the burden extension methods, in general, outperformed the Burden method. Among these burden extension methods, Burden-V1 and Burden-V2 had comparable power under all scenarios; Burden-S, Burden-V1/V2 outperformed Burden-A when any proportion of heteroplasmic variants displayed different effect directions; and Burden-V1/V2 outperformed Burden S for most scenarios. For example, by Definition 1, when 25% of heteroplasmic variants

were causal, and 80% of these causal variants had the same effect direction, Burden had a low power ($=0.29$) while Burden-A ($=0.63$), Burden-S ($=0.76$), Burden-V1/V2 ($=0.85$) had much higher power (Figure 2.1).

For a continuous trait, if other conditions were held constant, SKAT outperformed all burden methods if 5% or less of heteroplasmic variants were causal in a region (Figure 2.1). Of note, the power was also low (<0.6) for SKAT if $< 25\%$ of heteroplasmic variants were causal. If the proportion of causal heteroplasmic variants increased to 25% or higher, all burden methods displayed comparable or higher power than SKAT. For example, when 50% of the heteroplasmic variants were causal and 50% of the causal variants had the same effect direction, SKAT had a power of 0.63, Burden-S had a power of 0.65, and Burden-V1/V2 had a power of 0.89.

For a binary trait, most burden tests had comparable or higher power than SKAT when the proportion of causal heteroplasmic variants was 25% or higher (Figure 2.1), regardless of the effect direction. For example, Burden-V1 exhibited 156% greater power than the SKAT (0.41 versus 0.16) when 50% of the heteroplasmic variants were causal and 50% of these causal heteroplasmic variants had the same effect directionality. When only 5% of the heteroplasmic variants were causal and 50% of them had the same effect direction, neither Burden-V1 nor SKAT had power (0.002 versus 0.003).

Two omnibus tests by Definition 1

SKAT-O test had comparable power to SKAT under all scenarios. When other conditions were held constant, ACAT-O had a similar power to the more powerful gene-based test (i.e., a SKAT or Burden depending on the different scenarios), and therefore, ACAT-O was more powerful than SKAT-O when the real disease model was unknown. SKAT-O and ACAT-O displayed comparable power when 50% of the causal mutations had the opposite effect direction. However, ACAT-O was more powerful than SKAT-O if 80% or 100% of the causal mutations had the same effect direction (Figure 2.1 & Supplementary Figure 2.2).

Definition 2

By Definition 2, for both continuous and binary traits, SKAT outperformed all Burden tests when 5% of heteroplasmic variants were causal and other conditions were fixed. For example, when 5% of the heteroplasmic variants were causal and 80% of these causal mutations had the same effect direction, the power of SKAT and Burden-V1 were 0.91 and 0.53, respectively, with a continuous trait. These two methods had comparable power if 25% or more heteroplasmic variants were causal. For example, these two methods had a power of 0.92 when the proportion of causal heteroplasmic variants increased to 50% given other conditions were fixed. For the omnibus tests, SKAT-O had a

great power loss for both continuous and binary traits with Definition 2. For example, when 25% of the mutations were causal and 80% of the causal mutations had the same effect direction SKAT-O had a power of 0.15 while ACAT-O had a power of 0.56 for a binary trait (Figure 2.1, Supplementary Figures 2.1). The empirical power with and without empirical type I error adjustment was comparable (Figure 2.1, Supplementary Figures 2.1-2.3).

2.3.3 Application to Real Data

We identified heteroplasmic variants and performed quality control procedures in five TOPMed cohorts containing middle-aged and older participants [5456 African Americans (AA, mean age 59.2, women 60.8%) and 12,051 European Americans (EA, mean age 63, women 53.7%)] (Table 2.2, Supplementary Tables 2.3-2.4, Supplementary Figure 2.4). We then performed cohort- and ancestry-specific association analyses of heteroplasmic variants with age and sex. Covariates included batch variables or white blood cell count and platelet count. Meta-analysis was used to combine results in participants of European ancestry and those of African ancestry, and all participants of both ancestries. We reported meta-analysis results in all participants of both ancestries and compared results between ancestries.

Association of heteroplasmy with age

Two definitions of heteroplasmy coding tended to yield consistent p-value across methods in association testing with age in ancestry-specific meta-analyses and meta-analyses of all participants (Supplementary Tables 2.5-2.16). In a meta-analysis of all participants by Fisher's method, *RNR1*, *RNR2*, *CO1*, *CO2*, and *ND4* showed significant associations with age ($p < 0.001$) using either definition 1 or definition 2 by multiple methods (Table 2.3, Supplementary Tables 2.13-2.14). Using the fixed effect inverse variance method of the original burden test, *D-loop*, *RNR1*, *RNR2*, *CO1*, *CO3*, and *ND4*, *ND5*, *CYB* showed significant associations with age ($p < 0.001$) (Supplementary Tables 2.15-2.16). Using *RNR1* as an example, an increase by one heteroplasmy in this gene was significantly associated with 1.1 years of older age ($p = 3.8E-7$) by definition 1 (Supplementary Table 2.15). In addition, an increase by 1 SD in heteroplasmy VAF in *RNR1* was significantly associated with 0.036 years of older age ($p = 3.7E-11$) by definition 2 (Supplementary Table 2.16). Sensitivity analysis showed that association strength remained consistent after adjusting for white blood cell count, component counts, and platelet counts (Supplementary Figure 2.5).

Due to a large difference in sample sizes between participants of European ancestry and African ancestry, the results in the meta-analysis of all participants were more consistent with those in the meta-analysis of European participants (Supplementary Tables 2.5-2.16). In the meta-analysis of African

participants by Fisher's method for an association with age, *RNR2* was the only gene showing significant association ($p < 0.001$) by multiple burden tests but not by SKAT (Supplementary Tables 2.5-2.6). In the meta-analysis of EA participants, multiple genes, including *RNR1*, *RNR2*, *CO1*, *CO2*, and *ND4*, showed significant associations with age ($p < 0.001$) by multiple tests (Supplementary Tables 2.9-2.10).

Association of heteroplasmy with sex

Unlike the findings in association and meta-analysis with age, heteroplasmy in most mtDNA genes showed no association with sex (Supplementary Tables 2.17-2.28). For example, in the meta-analysis of all participants, *ND6* was the only gene associated with sex ($p < 0.001$) with two burden methods by definition 2 (Supplementary Table 2.26). In the meta-analysis of participants of African origin, *ND5* showed evidence of association with sex by multiple burden tests ($p < 0.001$) by definition 2 (Supplementary Tables 2.18 & 2.20). In meta-analysis with participants of European origin, no genes showed significant associations with sex (all $p > 0.001$) (Supplementary Tables 2.21-2.24).

2.4 Discussion

We proposed a framework that incorporates a pre-specified threshold for identifying true heteroplasmic variants and several gene-based tests to perform association analyses between heteroplasmic variants and a trait. We used simulation studies to evaluate the proposed framework in association analyses of mtDNA heteroplasmic variants and applied this framework to analyze age and sex with rare heteroplasmic variants in five large TOPMed cohorts with WGS.

Simulations studies

The proposed framework incorporates several gene-based methods and omnibus tests to provide a comprehensive evaluation of the trait-heteroplasmy association. The burden-extension tests outperformed SKAT for all simulation scenarios except for extremely unfavorable situations in which a very small proportion ($\leq 5\%$) of the heteroplasmic variants were causal and/or half of these causal variants display opposite directions. Under such unfavorable situations, the original burden had almost no power while the burden-extension tests had comparable power to SKAT. The original burden showed comparable power to burden extension methods only when $\sim 100\%$ of heteroplasmic variants showed consistent effect direction. Of the two omnibus tests, ACAT-O easily combines a large number of test p-values and it was more powerful than SKAT-O for most situations when combining SKAT and the original burden test.

It is also worth noting that our study showed the widely used weights, i.e., $\beta(\text{MAF}, 1, 25)$, and $\beta(\text{MAF}, 1, 1)$ provided similar results in association testing of heteroplasmy. While these methods outperformed the original burden test, the burden-extension tests provided only p-values without computing the effect size for a gene. The burden-extension tests use permutation to derive p-values, which is computationally extensive. While they are feasible in analyzing a small number of genes in mtDNA, these extension methods are challenging to analyze a large number of genes in nuclear DNA. The incorporation of several burden methods provides valuable underlying information about the proportion of heteroplasmic variants associated with the trait and their effect directions in the gene-based tests. We are currently extending the framework to identify these trait-associated heteroplasmic variants and classify the trait-associated heteroplasmic variants into distinct groups with different effect directions.

Association studies of age in real data

Multiple studies have reported that heteroplasmy is associated with somatic aging.^{52; 66} However, few studies have investigated gene-specific heteroplasmic somatic aging. In this study, we found that mtDNA displays different somatic aging rates among the 16 mtDNA encoded genes/area concerning the number of heteroplasmic variants and/or VAFs of variants. The strongest associations were found with the *RNR1* and *RNR2* genes. The fact that the Burden test yielded the most significant results compared to its extension

methods and SKAT by two definitions indicates that both the number of heteroplasmic variants and mutant allele fractions increase with advancing age. The *RNR1* and *RNR2* genes encode a 12S rRNA and 16S rRNA, respectively, and they are part of the machinery for the synthesis of 13 mtDNA-encoded polypeptides that are essential components of the mitochondrial oxidative phosphorylation (OXPHOS) pathway.⁶⁷ Despite their key roles in mitochondrial function, these two genes have been studied in far less detail than protein-coding genes in mtDNA. Mutations in *RNR1* were found to cause hearing loss.^{68; 69} More recently, a small open reading frame within *RNR2* that encodes the human polypeptide has been the target of Alzheimer's disease research.^{70; 71} Given that aging is the leading cause of Alzheimer's disease, heteroplasmic variants in these two genes merit further investigations for their relationships with Alzheimer's disease and other age-related diseases. The three cytochrome c oxidase (CO) subunits, *CO1*, *CO2*, and *CO3*, also showed significant association with age by more than one gene-based test with both definitions. Cytochrome c oxidase is the terminal complex (Complex IV) of the OXPHOS respiratory chain for aerobic metabolism.⁷² Maternally inherited mutations in the CO subunits are associated with many severe, inherited mitochondrial diseases, for example, Leber's hereditary optic neuropathy and complex IV deficiency.⁷³⁻⁷⁵

The *D-loop* region is the main non-coding, control area and a hot spot for mtDNA alterations. However, we did not observe a significant association ($p <$

0.001) between age and heteroplasmy in this region in either European Americans or African Americans. Our results support a previous finding that heteroplasmy in the mtDNA control region seems not to be the result of somatic age-related accumulation.⁷⁶ Of all coding genes, complex I consist of seven NADH-ubiquinone oxidoreductase (ND) genes encoded by mtDNA.⁷⁷ This complex is the first and the largest complex of the electron transport chain.⁷⁷ This complex oxidizes nicotinamide adenine dinucleotide (NADH) to generate electrons from NADH to coenzyme Q10 (CoQ10) and translocate protons across the inner mitochondrial membrane for energy metabolism.⁷⁷ Heteroplasmic variants in six of the seven subunits of complex I (except for *ND4*) showed no significant associations with advancing age by both definitions. In addition, heteroplasmic variants in the cytochrome b (*CYB*) gene and the two adenosine triphosphate (ATP) synthases were not significantly associated with age either. *CYB* is a component of respiratory chain complex III. This component is also involved in electron transport and the generation of proton gradient for the formation of the energy storage molecule ATP by the two enzymes, *ATP6* and *ATP8*, of the last complex (complex V) of the OXPHOS pathway.⁷⁸ These observations indicate certain cellular protection mechanisms may help prevent age-related increases in the number and frequency of heteroplasmy in these complexes. Over the years, various clinical phenotypes, mostly maternally inherited and severe, have been linked to mutations in the subunits of the OXPHOS complexes.^{77; 79}

In summary, the proposed framework provides a comprehensive evaluation of the trait-heteroplasmy association. Using this framework, we found that heteroplasmic variants are not likely to differ between men and women. We also found that somatic aging occurs unevenly across mtDNA regions, which merits further investigation between these identified age-related genes with age-related traits. This framework will facilitate association analyses of heteroplasmic variants with complex, age-related traits in large population data with WGS.

2.5 Tables and Figures

Table 2.1 Gene-wide empirical type 1 error rates with 95% confidence interval for coding definition 1 in simulation studies at $\alpha=0.001$

	Continuous Traits	Binary Traits (prevalence=20%)
Burden	0.88 (0.64, 1.18)	0.94 (0.69, 1.25)
A-Burden	1.22 (0.93, 1.57)	0.66 (0.45, 0.93)
Burden-S	1.06 (0.79, 1.39)	0.80 (0.57, 1.09)
Burden-V1	1.24 (0.95, 1.59)	0.32 (0.18, 0.52)
Burden-V2	1.56 (1.23, 1.95)	0.32 (0.18, 0.52)
SKAT	0.64 (0.44, 0.9)	1.20 (0.92, 1.54)
SKAT-O	0.64 (0.44, 0.9)	1.14 (0.86, 1.48)
ACAT	0.72 (0.5, 1)	1.18 (0.9, 1.52)

The number in each cell represents the ratio of type I error and the expected significance level of 0.001. Burden, original burden test; Burden-A, adaptive burden test; Burden-S, z-score weighting burden test; Burden-V1, variable threshold burden test with minimum p-value; Burden-V2, variable threshold burden test with ACAT p-value combination method; SKAT, sequence kernel association test; SKAT-O, sequence kernel association test-optimal test; ACAT, aggregated Cauchy association test combining burden and SKAT. We simulated 50,000 replicates for evaluating the type I error rate. We simulated a continuous variable and a binary variable in response to heteroplasmic variants located in the mitochondrial cytochrome b (MT-CYB) gene in European American participants (N=3,415) of the Atherosclerosis Risk in Communities (ARIC) Study.

Table 2.2 Participant characteristics in the five population-level cohorts with whole genome sequencing

Cohort	Sample size	Age, mean (\pm SD)	Female, n (%)	N of heteroplasmy
African American (n = 5456, women 60.8%)				
ARIC	241	58.4(6.3)	144 (59.8%)	162
CHS	705	73.8(5.6)	445 (63.1%)	673
JHS	3404	55.7(12.8)	2140 (62.9%)	1590
MESA	1106	60.9(9.6)	587 (53.1%)	968
European American (n = 12,051, women 53.7%)				
ARIC	3415	58.2(5.9)	1734 (50.9%)	1501
CHS	2788	74.2(5.7)	1594 (57.2%)	1859
FHS	3992	59.9(15.7)	2190 (54.9%)	2158
MESA	1856	61.5(9.8)	949 (51.1%)	1236

Total n, the total number of heteroplasmic sites identified in a cohort. ARIC, Atherosclerosis Risk in Communities Study; CHS, Cardiovascular Health Study; FHS, Framingham Heart Study; JHS, Jackson Heart Study; MESA, Multi-Ethnic Study of Atherosclerosis.

Table 2.3 Genes showing significant associations with age in the meta-analysis using Fisher's method

mtDNA region	P-values							
	Burden	Burden-A	Burden-S	Burden-V1	Burden-V2	SKAT	SKAT-O	ACAT
Definition 1								
MT-RNR1	1.07E-08	1.27E-08	4.45E-08	0.00013	0.0043	0.0075	6.06E-05	3.19E-08
MT-RNR2	2.23E-10	2.18E-10	2.17E-09	0.00045	5.11E-08	0.0073	1.49E-06	9.51E-09
MT-CO1	7.61E-05	1.10E-05	2.01E-05	0.057	0.00060	0.072	0.00039	0.00035
Definition 2								
MT-RNR1	5.36E-12	1.06E-08	4.19E-08	8.85E-05	3.19E-06	0.0015	0.0024	5.97E-11
MT-RNR2	5.52E-12	2.21E-10	2.24E-09	0.00041	4.17E-08	0.013	0.00093	2.06E-10
MT-CO1	2.06E-06	3.28E-06	4.56E-06	0.042	5.85E-05	0.050	0.019	2.16E-05
MT-CO2	0.011	0.025	0.023	0.053	0.042	0.0060	0.00070	0.0014
MT-ND4	0.0043	0.0044	0.0011	0.091	0.099	0.19	0.23	0.014

Cohort-specific association analysis was performed between heteroplasmic variants and age using gene-based tests and omnibus tests. Fisher's method was used to combine the p-values from individual cohorts/ancestries (i.e., meta-analysis) from all participants. This table includes genes that yielded $p \leq 0.001$ in any gene-based tests after meta-analysis. Burden, original burden test; Burden-A, adaptive burden test; Burden-S, z-score weighting burden test; Burden-V1, variable threshold burden test with minimum p-value; Burden-V2, variable threshold burden test with ACAT p-value combination method; SKAT, sequence kernel association test; SKAT-O, sequence kernel association test-optimal test; ACAT, aggregated Cauchy association test combining burden and SKAT.

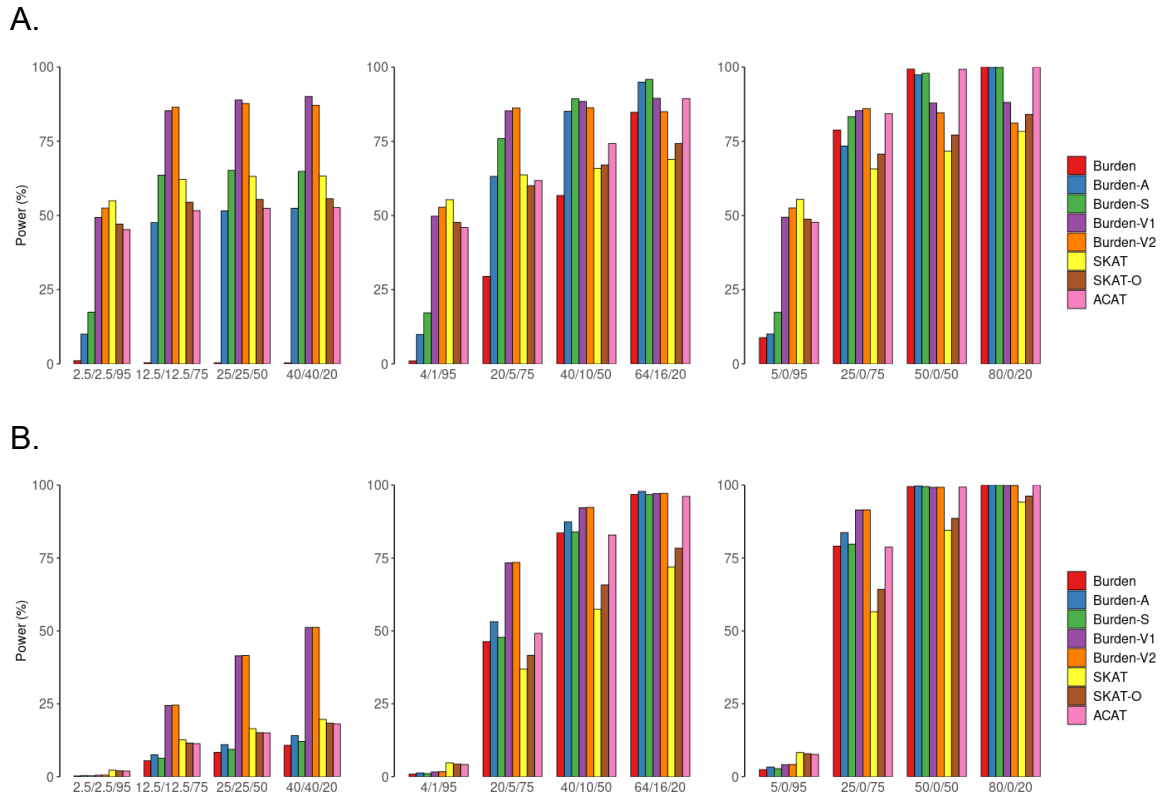


Figure 2.1 Simulation-based power comparisons of six gene-based tests and two omnibus tests with a continuous and a binary trait for coding definition 1 (adjusted for empirical type I error rate. Power estimation for a continuous trait (A) and a binary trait (B) at $\alpha=0.001$. Heteroplasmic variants are defined by an indicator function (definition 1). In simulations, we consider 5%, 25%, 50%, or 80% of the nonsynonymous heteroplasmic variants in the CYB gene to be causal and consider that 50%, 80%, and 100% of the causal heteroplasmic variants have effects with the same directionality. The variance that was explained by causal mutations was set to be 1% for the continuous trait and 2% for the binary trait. Burden, original burden test; Burden-A, adaptive burden test; Burden-S, z-score weighting burden test; Burden-V1, variable threshold burden test with minimum p-value; Burden-V2, variable threshold burden test with ACAT p-value combination method; SKAT, sequence kernel association test; SKAT-O, sequence kernel association test-optimal test; ACAT, aggregated Cauchy association test combining burden and SKAT. We simulated 50,000 replicates for evaluating power.

CHAPTER 3 CLUSTERING OF RARE VARIANTS FOR CAUSAL VARIANTS IDENTIFICATION AND EFFECT DIRECTION CLASSIFICATION

3.1 Introduction

During the last decade, genome-wide association studies (GWASs) have identified hundreds of thousands of common genetic variants (minor allele frequency $\geq 5\%$) associated with numerous complex diseases and quantitative traits.^{1; 80} In addition, low-frequency variants ($1\% \leq \text{MAF} < 5\%$) and rare variants ($\text{MAF} < 1\%$) are detected increasingly with the advent of next-generation sequencing technologies. Low-frequency variants and rare variants are substantial sources of unexplained heritability of various phenotypes.⁴⁵ Several gene-based tests such as the burden test⁴⁶ and sequence kernel association test (SKAT)⁴⁷ have been developed for association testing of rare single nucleotide variants (SNVs) in genomic regions with disease traits. It has been shown that none of these gene-based tests is uniformly most powerful. Hence, omnibus tests such as SKAT-O⁴⁸ and aggregated Cauchy association test (ACAT-O)⁴⁹ are proposed to search for an optimal combination of the burden test and SKAT to provide robust summary statistics. Unlike single/multiple variant models, a common limitation of these aggregate methods is that they do not discriminate potential causal variants from null variants in association testing within the tested regions.

To overcome this weakness, we propose a clustering method based on a Gaussian mixture model (GMM) to discriminate potentially causal rare variants from null variants in the gene regions that are associated with disease traits. In large GWAS and meta-analysis, gene-based association analyses (e.g., SKAT-O and ACAT-O) are conducted to combine the burden test and SKAT to detect significant genes or regions (deemed as “signal regions” in this dissertation) in which rare variants show associations with a trait. The burden test outperforms SKAT when a large proportion of rare variants within a region are trait-associated, and most of the trait-associated rare variants have the same effect direction. The SKAT has a larger power than the burden test otherwise.⁴⁷ For a given signal region associated with a disease trait, we fit multiple-variant models to obtain association statistics between phenotype and rare variants within the region. Based on the variant-level statistics, this novel clustering method may identify potentially risk and/or protective genetic variants in a genomic region. Furthermore, the method may also cluster the signal variants into subgroups of variants with different effect sizes and effect directions in association testing. We simulate genomic regions with independent rare variants and variants in linkage disequilibrium (LD). We evaluate the performance of the proposed method by a comprehensive simulation study with several statistics, including the adjusted rand index (ARI), mean square error (MSE), and accuracy of the number of clusters specification. We then apply the new clustering method to identify risk and protective rare variants in six genes that are significantly associated with

blood pressure (BP) traits in the most recent large GWAS and meta-analysis.⁸¹

The identification of risk and protective rare variant clusters is critical not only for investigating the underlying biological mechanism between rare variants and disease traits, but also for the identification of drug targets and the design of gene therapy.

3.2 Methods

3.2.1 Association testing of rare variants

The burden test⁴⁶ collapses the rare variants within a gene region into a single genetic score and then tests the association of that variable with the trait. SKAT⁴⁷ employs a score-type variance component test. The test statistics of these two tests can be constructed by the score statistics of rare variants in the tested region. The score statistics of j^{th} rare variant are given by

$$U_j = \sum_{i=1}^n G_{ij}(y_i - \hat{\mu}_i)$$

where G_{ij} is the genotype or genetic dosage of the i^{th} individual at the j^{th} locus, $\hat{\mu}_i$ is the estimated mean of the outcome y_i under the null hypothesis that there is no association between the variants in the gene region and the phenotype. The test statistic of the burden test is $Q_{burden} = (\sum_{j=1}^J w_j U_j)^2$, where J is the number of rare variants within the target region and w_j is the weight of the j^{th} variant.

Under the null, Q_{burden} follows an asymptotic chi-square distribution with 1

degree of freedom.

The burden test is most powerful when all rare variants in the target region have the same effect direction and similar effect sizes. The burden test is underpowered when a large proportion of rare variants are non-causal or have opposite effect direction. The SKAT outperforms the burden under such scenarios. The SKAT test statistic is $Q_{SKAT} = \sum_{j=1}^J w_j^2 U_j^2$ which follows a mixture of independent chi-square distributions asymptotically with 1 degree of freedom under the null. To improve power by combining these two tests, we adopt the ACAT-O method⁴⁹ with two commonly used choices of weights based on MAF of beta density. The test statistic of the ACAT is $T_{ACAT-O} = \sum_g \sum_b \tan\{(0.5 - p_{g,b})\pi\}$ where $g \in \{burden, SKAT\}$ $b \in \{beta(1,1), beta(1,25)\}$. The p-value of ACAT is calculated by $p_{ACAT-O} \approx \frac{1}{2} - \frac{\arctan(T_{ACAT-O}/4)}{\pi}$. The clustering analysis of rare variants is performed only if the p-value of ACAT-O is smaller than a selected threshold, i.e., $p_{ACAT-O} \leq \alpha$, for a given region.

3.2.2 Multiple-variant model to obtain variant level summary statistics

If a gene region shows statistical significance with a disease trait ($p_{ACAT-O} \leq \alpha$), we perform multiple-variant analysis for all rare variants within the region to account for potential LD between the rare variants. By adjusting for covariates, the multiple-variant model is given by

$$g(\mu_i) = \alpha_0 + X_i^T \alpha + \sum_j^J G_{ij} \beta_j$$

where $g(\cdot)$ is the identity function for a continuous trait and the logistic function for a binary trait. X_i^T is a row vector of covariates of the i^{th} individual, G_{ij} is the genotype of this individual at the j^{th} locus within the target gene region. The beta coefficient $\hat{\beta}_j$, standard error $SE(\hat{\beta}_j)$ and corresponding p-value p_j of the j^{th} rare variant are obtained from the multiple-variant model. The beta coefficient and standard error statistics are used for the subsequent clustering analysis.

3.2.3 Gaussian mixture model

The Gaussian mixture model assumes data is generated from a mixture of Gaussian (normal) distributions. We assume that each beta coefficient from a multiple-variant model has its variance which is estimated by $\hat{\sigma}_j = SE(\hat{\beta}_j)$. We also assume that each beta coefficient can be classified into two types of clusters: a null cluster and K signal clusters. Therefore, the total number of clusters is $K^* = K + 1$. A null cluster includes all non-causal rare variants of a gene region. Within each of the K signal clusters, the causal variants have a similar effect on the trait. With a null cluster and K signal clusters, each beta coefficient β_j of the multiple-variant model follows one of the normal distributions of $N(\mu_0 = 0, \hat{\sigma}_j^2), N(\mu_1, \hat{\sigma}_j^2), \dots, N(\mu_K, \hat{\sigma}_j^2)$, where μ_k is the true mean of beta estimates for rare variants classified in the same cluster. $\hat{\sigma}_j^2$ is the estimated

variance of the j^{th} beta coefficient from the multiple-variant model, $\mu_0 = 0$ is the mean of the null cluster. For each β_j , we introduce a hidden (unobserved) class label variable D_j which follow a categorical distribution (that is the generalized Bernoulli distribution). The class label variable D_j indicates the component (cluster) that $\hat{\beta}_j$ belongs to; that is,

$$p(\hat{\beta}_j | D_j = k, \mu_k) = \frac{1}{\hat{\sigma}_j^2 \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{\hat{\beta}_j - \mu_k}{\hat{\sigma}_j} \right)^2}$$

$$p(D_j = k) = \varphi_k, \quad 0 \leq \varphi_k \leq 1 \text{ and } \sum_{k=0}^K \varphi_k = 1$$

φ_k is the proportion of component k in the mixture distributions. The density function of each data point $\hat{\beta}_j$ is given by

$$p(\hat{\beta}_j | \boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^K p(\hat{\beta}_j | D_j = k, \mu_k) \varphi_k$$

where $\boldsymbol{\mu} = \{\mu_0, \dots, \mu_K\}$ and $\boldsymbol{\varphi} = \{\varphi_0, \dots, \varphi_K\}$. The log-likelihood function of all beta coefficients $\hat{\boldsymbol{\beta}} = \{\hat{\beta}_1, \dots, \hat{\beta}_j\}$ is given by

$$l(\boldsymbol{\theta}, \boldsymbol{\pi} | \mathbf{X}) = \log \prod_{i=1}^n \sum_{k=1}^K p(\hat{\beta}_j | D_j = k) \varphi_k = \sum_{i=1}^n \log \sum_{k=1}^K p(\hat{\beta}_j | D_j = k) \varphi_k$$

Because there is no closed-form solution of the maximum likelihood estimators (MLE) of the parameters, the expectation–maximization (EM) algorithm is employed to find a numeric solution. The EM algorithm is an iterative method that is described in detail in the next section.

3.2.4 Parameter estimation by expectation maximization (EM) algorithm

Given a fixed number of signal clusters K , we employ the EM algorithm to estimate the $2K$ parameters $\{(\mu_k, \varphi_k) | k = 1, \dots, K\}$ because φ_k are sum to 1, that is, $\sum_{k=0}^K \varphi_k = 1$. The algorithm undergoes three steps: (1) an initialization step to set the initial values of the parameters; (2) an expectation step that computes the expected value of the log-likelihood; and (3) a maximization step to update the parameters by maximizing the expectation of the log-likelihood. (4) determine the number of clusters K^* by the minimum BIC with $2 \leq K^* \leq 7$.

Initialization

The EM algorithm may converge to a local optimum, and the chance of converging to a global optimum (MLE) depends essentially on the initial values of the parameters.⁸² To minimize the effect of initialization, we set initial values by a crude estimation of the proportion of non-causal variants φ_0 using a variable threshold approach. Suppose the p-values of multiple-variant models of all the rare variants within a region are $S = \{p_j, j = 1 \dots J\}$. We define an absolute value of the standardized beta coefficient (Z-score) as $|Z_j| = \frac{|\hat{\beta}_j|}{SE(\hat{\beta}_j)}, j = 1, \dots, J$. For each p-value, we use a variable threshold approach, that is, p_l is used a threshold, here, $p_l \in S$. We meta-analyze the $|Z_j|$ whose corresponding p-value $p_j \leq p_l$ by the weighted sum of the Z-scores method: $\Omega_l = \frac{\sum_{j:p_j \leq p_l} a_j |Z_j|}{\sqrt{\sum_{j:p_j \leq p_l} a_j^2}}$, where a_j is a pre-

specified weight for j^{th} absolute value of standardized beta.

Different weighting schemes can be used. An appropriate choice of weight may improve the accuracy of the estimation of φ_0 . We select two weighting schemes: equal weight ($a_j = 1$) and weight as the inverse of the estimated standard error ($a_j = \frac{1}{SE(\hat{\beta}_j)}$). Given a specified weighting scheme, we define an optimal threshold as $p_0^* = \arg \max_{p_l} \Omega_l$. Then, we define a rare variant as a preliminary signal variant if the corresponding p-value $p_j \leq p_0^*$; otherwise, the variant is defined as a preliminary non-signal variant. The initial value of φ_0 can be calculated by $\widehat{\varphi}_0^{(0)} = 1 - \frac{\sum_j I_{\{p_j \leq p_0^*\}}}{|S|}$. $|S|$ is the number of p-values in the set. The preliminary signal variants are then clustered by K-means analysis. The remaining mixing proportion parameters $\{\widehat{\varphi}_1^{(0)}, \dots, \widehat{\varphi}_K^{(0)}\}$ are calculated as $1 - \widehat{\varphi}_0^{(0)}$ multiplied by the proportion of variants in the K signal clusters from the K-means clustering. The initial values of the means of the signal clusters $\{\widehat{\mu}_1^{(0)}, \dots, \widehat{\mu}_K^{(0)}\}$ are set as the cluster means from the K-means analysis.

We add a constraint that the preliminary signal variants with opposite effect directions cannot be assigned to the same cluster by the K-means method. That is, the variants with positive beta coefficients and negative ones are split into two separate clusters by the K-means method. Specifically, $K = K_+ + K_-$, where K_+ and K_- are defined as the number of clusters for the variants with

positive and negative beta coefficients, respectively. Then we run K-1 combinations $((K_+ = 1, K_- = K - 1), \dots, (K_+ = K - 1, K_- = 1))$ of K-means analysis. The optimal cluster partition is identified by minimizing the within-cluster sum of squares (WCSS) across the K-1 combinations:

$$\arg \min_{S^+, S^-, K_+} \left(\sum_{k_+=1}^{K_+} \sum_{\beta \in S_{k_+}} \|\beta - \mathbf{c}_i\|^2 + \sum_{k_-=1}^{K-K_+} \sum_{\beta \in S_{k_-}} \|\beta - \mathbf{c}_i\|^2 \right)$$

Where $S^+ = \{S_1^+, \dots, S_{K_+}^+\}$ is a partition of the variants with positive effect direction, and $S^- = \{S_1^-, \dots, S_{K_-}^-\}$ is a partition of variants with a negative effect direction.

Expectation step

Because the EM algorithm is an iterative method, we designate $\hat{\boldsymbol{\mu}}^{(i)}$ and $\hat{\boldsymbol{\varphi}}^{(i)}$ as the values of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\varphi}}$ for i^{th} iteration, respectively. For simplicity, we use the $\hat{\sigma}_j^2$ value estimated from the multiple-variant models, and therefore, we do not update $\hat{\sigma}_j^2$ in the EM algorithm. Then we define the posterior probabilities of D_j by $\hat{\boldsymbol{\mu}}^{(i)}$ and $\hat{\boldsymbol{\varphi}}^{(i)}$ for each iteration

$$P_i(D_j = k | \hat{\boldsymbol{\beta}}_j) = \gamma_{ij}(k) = \frac{p(\hat{\boldsymbol{\beta}}_j | D_j = k, \hat{\boldsymbol{\mu}}_k^{(i)}) \hat{\varphi}_k^{(i)}}{p(\hat{\boldsymbol{\beta}}_j | \hat{\boldsymbol{\mu}}^{(i)}, \hat{\boldsymbol{\varphi}}^{(i)})}$$

The expectation of the log-likelihood w.r.t current posterior probabilities of D_j given $\hat{\boldsymbol{\beta}}$ is

$$E_{D|\widehat{\beta}, \widehat{\mu}^{(i)}, \widehat{\varphi}^{(i)}} l(\boldsymbol{\theta}, \boldsymbol{\varphi} | \mathbf{X}) = \sum_{j=1}^J \sum_{k=0}^K \gamma_{ij}(k) \log(p(\widehat{\beta}_j | D_j = k, \mu_k) \varphi_k)$$

Note that $\mu_0 \equiv \widehat{\mu}_0^{(i)} \equiv 0$.

Maximization step

To maximize parameters, we take partial derivatives of $E_{D|\widehat{\beta}, \widehat{\mu}^{(i)}, \widehat{\varphi}^{(i)}} l(\boldsymbol{\theta}, \boldsymbol{\varphi} | \mathbf{X})$ w.r.t each parameter of $\boldsymbol{\mu} = \{\mu_1, \dots, \mu_K\}$ and $\boldsymbol{\varphi} = \{\varphi_1, \dots, \varphi_K\}$ and set to 0. That is

$$\sum_{j=1}^J \gamma_{ij}(k) \frac{\partial}{\partial \mu_k} \log(p(\widehat{\beta}_j | D_j = k, \mu_k) \varphi_k) = 0, k = 1, \dots, K$$

$$\sum_{j=1}^J \gamma_{ij}(k) \frac{\partial}{\partial \varphi_k} \log(p(\widehat{\beta}_j | D_j = k, \mu_k) \varphi_k) = 0, k = 1, \dots, K$$

By solving the above equations and adding the constraint of $\sum_{k=0}^K \varphi_k = 1$, we have

$$\widehat{\mu}_k^{(i+1)} = \frac{\sum_{j=1}^J \gamma_{ij}(k) \widehat{\beta}_j / \sigma_j^2}{\sum_{j=1}^J \gamma_{ij}(k) / \sigma_j^2}$$

$$\widehat{\varphi}_k^{(i+1)} = \frac{\sum_{j=1}^J \gamma_{ij}(k)}{J}$$

The expectation and maximization steps are implemented iteratively. The

algorithm stops when $|l(\boldsymbol{\theta}, \boldsymbol{\pi}|\mathbf{X})^{(i+1)} - l(\boldsymbol{\theta}, \boldsymbol{\pi}|\mathbf{X})^{(i)}| \leq \varepsilon$ for with a small positive number ε . We set $\varepsilon = 0.001$.

For a given K^* , we choose the clustering results with a higher likelihood from the two sets of initialization values. We adopt the Bayesian information criterion (BIC) to determine the optimal number of clusters with $2 \leq K^* \leq 7$: $BIC = (2K) \ln(J) - 2 \ln(\hat{L})$, where J is the number of variants and \hat{L} is the maximized value of the likelihood function.

3.2.5 A simulation study

We simulated a continuous phenotype by the following model:

$$y = 10 + 0.6X_1 + 0.8X_2 + \mathbf{G}_C^T \boldsymbol{\beta} + \varepsilon$$

where 10 is the intercept, $X_1 \sim N(0,1)$, $X_2 \sim \text{Binomial}(0.5)$ and $\varepsilon \sim N(0,0.49)$. $\mathbf{G}_C^T = (G_{C1}, \dots, G_{CL})$ is a vector that includes the genetic coding for L randomly chosen causal rare variants in the simulated region. $\boldsymbol{\beta} = (\beta_1, \dots, \beta_L)^T$ is a vector of true beta effects for the selected causal variants. The beta effect of rare variant i is based on R^2 and $\boldsymbol{\omega}$. R^2 is the proportion of variance explained by all of the causal rare variants for a continuous trait. $\boldsymbol{\omega} = \{\omega_1, \dots, \omega_L\}$ is the proportion of the beta effects of each causal rare variant. The elements of $\boldsymbol{\omega}$ are randomly assigned to

each causal rare variant. $\beta_l = \text{sign}(\omega_l) \sqrt{\frac{cs_l^2}{2MAF_l(1-MAF_l)}}$, $c = \frac{R^2}{s^T D s}$, where D is the correlation matrix between causal variants, and $s = (s_1, \dots, s_l)$, where $s_l = \text{sign}(\omega_l) \sqrt{2MAF_l(1-MAF_l)\omega_l^2}$. The proportion of variance (R^2) explained by the causal rare variants was set to be 3% for the continuous phenotype. We also applied a cutoff of 80% quantile to the simulated continuous phenotype with $R^2 = 3\%$ to obtain a binary phenotype. The variance explained by X_1 , X_2 and random error was around 36%, 16%, and 45% respectively.

We conduct a simulation study of 1,000 replicates to evaluate the performance of the proposed clustering method. We generate 60 rare variants with a minor allele frequency (MAF) in $0.002 \leq \text{MAF} < 0.01$. It has been shown that rare variants display mild linkage disequilibrium (LD) between each other.⁸³ LD may affect the performance of the clustering because the GMM model assumes the data points are independent. To evaluate the effect of LD on rare variant clustering, we perform simulations to generate rare variants under two conditions: independent variants and those that display LD using human genome sequence data from the 1000 Genome Project Build 37 as the reference.⁸⁴ For the first condition, we generate genotypes using the PhenotypeSimulator R package.⁸⁵ This package simulates genotypes based on the binomial distribution that does not incorporate LD. Second, we generate genotypes using the sim1000g R package that uses the first 60 rare variants of the B-Cell

Translocation Gene 3 (BTG3) gene (19,249 base pair in length, GRCh37/hg19) on chromosome 21 in the human 1000 Genomes Project (Build 37) as the reference. Of note, the choice of the BTG3 gene is arbitrary. The sim1000G package simulates variants for small or large genomic regions or a full chromosome in unrelated individuals or family data. Haplotypes are extracted to compute LD in the simulated genomic regions and to generate new genotype data among individuals. For both situations, i.e., the presence and absence of LD by the two R packages, we simulate 1000 replicates of 5,000, 15,000, 25,000 samples. To compare the effect of sample size on rare variant clustering, we randomly select 5000, 15,000 and 25,000 samples from the underlying population of 250,000 individuals. To understand the LD structure between rare variants generated by the sim1000g package, we calculate an average correlation matrix of the simulated 60 rare variants over the 1000 replicates.

For both the continuous and binary traits, we compare the performance of the clustering methods under six simulation scenarios (**Table 3.1**). For scenario 1: 1/3 of the variants within the target region is non-causal, 1/3 of the variants are causal with moderate positive effect, and 1/3 of the variants are causal with strong positive effect. The ratio of moderate and strong effects is 1:2. In scenario 2: 1/3 of the variants within the target region are non-causal, 1/3 of the variants are causal with positive effect, and 1/3 of the variants are causal with negative effect. The effect size of each causal variant is the same for this scenario. In

scenario 3: 1/4 of the variants within the target region are non-causal, 1/4 of the variants are causal with moderate positive effect, 1/4 of the variants are causal with strong positive effect, and 1/4 of the variants are causal with moderate negative effect. The ratio of the effects is -1:0:1:2 for this scenario. In scenario 4: 1/5 of the variants within the target region is non-causal, 1/5 of the variants are causal with moderate positive effect, 1/5 of the variants are causal with strong positive effect, 1/5 of the variants are causal with moderate negative effect, and 1/5 of the variants are causal with strong negative effect. The ratio of the effects is -2:-1:0:1:2. In scenario 5: 1/2 of the variants are causal with moderate positive effect, and the other 1/2 of the variants are causal with strong positive effect. The ratio of the effects is 1:2. In scenario 6: 1/2 of the variants are causal with positive effect, and the other 1/2 of the variants are causal with negative effect. The ratio of the effects is -1:1. Note that scenarios 5 and 6 are two extreme cases that do not contain a null cluster.

We adopt the Adjusted Rand index⁸⁶ (ARI) method to measure the similarity between the true and predicted allocations of clusters. That is, we calculate the average ARI over the 1000 replicates to evaluate the performance of the method. For each scenario, we also calculate the accuracy of the number of clusters K^* determination. The accuracy is calculated as

$$\text{Accuracy} = \frac{\text{\# of replicates determine } K^* \text{ correctly}}{\text{\# of replicates}}$$

To assess the deviation between mean effects in estimation and the “true” effects from the simulation from clusters, we calculate the mean squared error (MSE) by

$$\text{MSE} = \frac{\sum_r \sum_j (\beta_{rj} - \widetilde{\beta}_{rj})^2}{J \times \# \text{ of replicates}}$$

where β_{rj} is the true mean of the j^{th} cluster in the r^{th} replicate; and $\widetilde{\beta}_{rj}$ is the corresponding estimated mean based on our clustering algorithm. For both simulation studies and real data analyses, we perform an analysis of variance (ANOVA) to test if the means of betas from different clusters are significantly different. We also apply the Mann-Whitney U test to evaluate if the mean of betas from a null cluster is significantly different from 0.

3.2.6 Application to exome-wide association and a rare-variant GWAS of blood pressure traits

To demonstrate the proposed clustering method, we apply the method to cluster the rare variants within the significant genes associated with blood pressure (BP) traits. Blood pressure is an inherited trait with an estimated heritability of up to 30-70%.⁸⁷ High blood pressure is an independent risk factor for cardiovascular diseases.⁸⁸ We apply the proposed clustering method to summary statistics obtained from the most updated association studies and meta-analysis of BP traits by Surendran et al.⁸⁹ This study included more than 800,000 individuals from four consortia (CHARGE, CHD Exome+,

GoT2D:T2DGenes, ExomeBP) and UK BioBank data.⁸⁹ In this study, an exome-wide association (EWAS) and a rare-variant GWAS (RV-GWAS) using imputed and genotyped single nucleotide variants (SNVs) were conducted to identify common and rare variants, and genes that were associated with three continuous BP traits (systolic blood pressure [SBP], diastolic blood pressure [DBP] and pulse pressure [PP]) and hypertension (HTN) by using both single-variant model and gene-based tests. This large effort validated most of the previously identified BP-associated single variants and genes. In addition, this large study discovered several new SNVs and genes associated with BP traits.⁸⁹ To cluster rare variants, we consider rare variants in genes that are associated with SBP, DBP, PP, and HTN.⁸⁹ We apply two strategies to cluster rare variants. In the first strategy, we cluster rare variants per gene and trait. In the second strategy, we cluster the combined rare variants per gene.

3.3 Results

3.3.1 Simulation studies

We first compared clustering methods under different scenarios using independent rare variants generated by the PhenotypeSimulator R package.⁸⁵ We then evaluated the effect of LD on the clustering method for rare variants generated by the sim1000g R package.

Simulations without LD structure

The ARI value was largely improved when the sample size was increased, indicating that a large sample size provided a more accurate allocation of the clusters. For example, with a sample size of 5000, the mean ARI value was 0.61 for the combined weighting scheme for a continuous trait under simulation scenario 2. With a sample size of 15,000 and 25,000, the ARI value increased to 0.90 and 0.96, respectively. The ARI values were comparable using three Z-score weighting schemes under each scenario (Supplementary Figures 6 and 7). Simulation scenarios 2 and 6 had higher mean ARI values compared to the other scenarios if other conditions were the same. This observation was as expected because the two signal clusters had opposite effect directions and the differences between the true cluster means of the two signal clusters were larger than those of the other scenarios.

With a pre-specified number of clusters K^* ranging from 2 to 7, the accuracy of determining the number of clusters was high (>0.9) for all simulation scenarios when the sample size reached 25000 for a continuous trait. The MSE of the estimates for the true cluster means was reduced with the increase in sample size. The ANOVA test was significant for all replicates under each scenario ($P < 0.05$), indicating that the means of betas from the different clustering were significantly different. The Mann-Whitney test presented non-significant

results for all the scenarios ($P > 0.05$), indicating that the rare variants allocated to the null cluster display effect sizes not significantly different from 0.

The ARI value of clustering for binary outcomes was lower than that of continuous outcomes when all the other conditions remained the same (Supplementary Figure 3.2-3.13). For example, when there was no LD and the sample size was 15000, the mean ARI value of a continuous trait with the combined weighting was 0.67 under scenario 1. The corresponding mean ARI value of a binary trait was 0.34.

Comparison of Simulations with and without LD

We observed a moderate or low LD between the 60 simulated rare variants using the sim1000g package (Supplementary Figure 3.1). Among a total of 1770 rare variant pairs from the 60 rare variants, 530 pairs (29.9%) displayed a correlation between 0.01 and 0.05, with 73 pairwise displaying correlations > 0.05 . The strongest correlation between the two variants was 0.26. The two variants had MAF 0.0031 and 0.0032.

For simulation scenarios 2, 4, and 6, the simulated genotype displaying mostly moderate LD provided comparable mean ARI values, MSE, and accuracy

of specification of the number of clusters K^* compared to those of independent rare variants (Figure 2, Supplementary Figure 4). For example, the mean ARI values with/without the presence of LD among variants were comparable (0.91 vs. 0.9) in scenario 2 with a sample size of 15,000 for a continuous trait using the combined weighting scheme. The corresponding MSE (0.0023 vs. 0.0022) and accuracy of specification of the number of clusters K^* (0.992 vs. 0.995) were similar between the presence and absence of LD structure. For simulation scenarios 1, 3, and 5, the simulated genotype displaying mostly moderate LD provided a lower mean ARI value, larger MSE, and lower accuracy of specification of the number of clusters K^* compared to those of independent rare variants (Table 3.2, Supplementary Tables 3.1–3.5, Supplementary Figures 3.2–3.5). For example, using the combined weighting scheme, in a sample size of 15,000 with a continuous trait, the average ARI value was 0.53 when the simulated rare variants displayed LD for scenario 1. In contrast, when simulated genotypes were independent, the average ARI value was 0.67, which was about 0.14 higher than the clustering result based on the variants with LD using the same sample size. The corresponding MSE from clustering rare variants without LD structure was 0.0033 (Table 3.2), while the MSE from cluster rare variants with LD was 0.0043 (Supplementary Table 3.1). The specification accuracy of the number of K^* clusters with the presence of LD (0.715) was lower than that (0.911) with the absence of LD between variants. Comparing the clustering performance of the scenarios between the presence and absence of LD, we observed that the

presence of LD had a larger effect on the performance of clustering for scenarios 1, 3, and 5 compared to scenarios 2, 4, and 6. It was likely due to the smaller differences between the true cluster means in scenarios 1, 3, and 5 compared to those in scenarios 2, 4, and 6.

3.3.2 Application to GWAS of BP traits with rare variants

Identification of blood pressure trait-associated genes

Using the SKAT test, multiple rare variants ($MAF < 0.01$) were identified for one or more BP traits ($P < 2.5 \times 10^{-6}$) with four genes (*NPR1*, *DBH*, *COL21A1*, and *NOX4*).^{81; 90} Low frequency and rare variants in two additional genes of *PLCB3* and *CEP120* were associated with BP traits at $MAF < 0.05$. The six genes harbor different numbers of rare variants. More specifically, *NPR1* included 13 rare variants, *DBH* included 29 rare variants, *COL21A1* included 26 rare variants and *NOX4* included 9 rare variants (Supplementary Table 3.6). SBP was associated with *NPR1*, *DBH*, and *PLCB3*; DBP was associated with *DBH* and *PLCB3*; PP was associated with *COL21A1*, *NOX4*, and *CEP120* due to multiple rare variants in the gene ($P < 2.5 \times 10^{-6}$). Because gene-based test results were not available for the associations between HTN and these six genes in the GWAS⁸⁹, we defined a signal gene of HTN if any of the six genes contained rare variant(s) ($MAF < 0.01$) with $P < 1 \times 10^{-4}$ in the single variant-HTN association testing. The three genes of *DBH*, *NPR1*, and *PLCB3* included rare variant(s) displaying association

with HTN at $P < 1E-4$. We applied the proposed clustering method to cluster BP-associated rare variants in these genes.

Clustering of rare variants

We performed rare variant clustering in eleven gene-trait associations (six genes with four traits): three genes associated with SBP, two genes associated with DBP, three genes associated with PP, and three genes that contain significant rare variants with HTN. On average, a signal gene contains about 20 rare variants. Two to three clusters were identified in each of the eleven gene-trait associations. About 70% of the rare variants were clustered into a null cluster (Tables 3.3-3.4, Supplementary Tables 3.7-3.24). For example, the NPR1 gene contained 12 rare variants. We identified three clusters of rare variants in this gene for the HTN association. The means of standardized beta coefficients (z-scores) were significantly different across the three clusters (ANOVA $P=0.00028$). Of the 12 rare variants, 8 variants were in the null cluster. The rare variants allocated to the null cluster displayed effect sizes (i.e., the standardized beta coefficients) not significantly different from zero (Mann-Whitney U test $P=0.46$). Three variants, including rs140425746, rs61757359, and rs61758562, were grouped into a cluster with an average effect size of -2.05. The rare variant, rs116245325, which displayed the smallest p-value of $1.46E-5$ with HTN in single variant analysis, forms a single cluster with an effect size of 4.33. Of note, the

clustering pattern in this example was similar to the layout in scenario 2 of the simulation study. That is, the two signal clusters had opposite effect directions. (Supplementary Table 3.23)

We also conducted clustering analyses on the combined summary data from significant genes that were associated with each of the BP traits. Three genes were associated with PP traits. (Supplementary Table 3.6) A total of 55 variants were located in these three genes. Of the 55 rare variants within these three genes, we identified five clusters with distinct effect sizes (ANOVA $P=1.39E-19$). (Table 3.3, Supplementary Tables 3.12) The null cluster contained 43 out of 55 variants. The effect sizes of rare variants allocated to the null cluster displayed were not significantly different from zero (Mann-Whitney U test $P=0.976$). Two rare variants, rs139341533 and rs56061986, were clustered together with an average effect size of -0.097. Four rare variants, rs2303720, rs114280473, rs189429890, and rs144215891, were assigned to a distinct cluster with an average effect size of -0.0438. A single variant, rs200999181, was recognized as the only variant in the cluster with the strongest effect size of 0.334. Five rare variants, rs201955087, rs115079907, rs76146749, rs200401514, and rs2764043, were grouped into a signal cluster with a moderate positive effect size of 0.173. The observed five-cluster pattern was similar to that in scenario 4 of the simulation study. Of note, scenario 4 included four signal clusters: a cluster with a strong positive effect, a cluster with a moderate positive

effect, a cluster with a strong negative effect, and a cluster with a moderate negative effect. For both positive and negative effects, the strong effect sizes were around twice the moderate effect sizes. The variants within the clusters of moderate effect sizes were more than the ones within the clusters of strong effects. The effect size of positive associations was larger than the effect size of negative ones. The running time of the BP variants clustering was around 33 seconds with a core of 4GB of memory.

3.4 Discussion

We proposed a new method to cluster rare variants within signal gene regions associated with disease traits based on summary statistics of variant-trait associations. We performed a comprehensive simulation study to evaluate the performance of the proposed method under different scenarios concerning variant effect direction, LD structure, the number of clusters, and the study sample size. In simulation scenarios 1, 3, and 5, we observed that the proposed method provided a higher ARI value, a lower MSE, and a higher accuracy in the specification of the number of clusters with rare variants in the absence of LD compared to those with low LD given the other conditions are the same. This is expected because the GMM model assumes that the data points (i.e., rare variants) are independent. The ARI value is improved and the MSE value is decreased with an increase in sample size. Among the simulated scenarios,

scenario 2 yielded the highest ARI value, smallest MSE, and highest accuracy in the specification K^* clusters because this scenario had the largest differences between the true means between clusters.

We also applied the proposed method to cluster rare variants of signal genes associated with BP traits using summary statistics from a large GWAS and meta-analysis of BP traits using around 1.3 million individuals.⁸¹ We first classified the rare variants in individual genes per trait. To further demonstrate the classification utility, we applied the methods with combined variants in several genes that are associated with the same traits. We found that most of the rare variants within the BP traits-associated genes are grouped into the null cluster, indicating that natural selection is likely the main force in shaping the rare variants in the human genome.⁹¹ In addition, we found that more rare variants were allocated to clusters with negative means than those with positive means.

A common limitation of gene-based tests is that these methods are not able to cluster possibly neutral, risk, and/or protective rare variants in trait-associated regions. The proposed method may overcome this obstacle by clustering rare variants based on the summary statistics at the single variant level obtained from the published large GWAS and meta-analyses (e.g., > 500,000 samples). By performing the proposed clustering analyses, we can distinguish

variants with opposite effects or those with different levels of effect sizes in the target region. In addition, the proposed method is computationally efficient for analyzing large-scale sequencing data.

Previous studies have reported inconsistent views about LD structure among rare variants. Some studies assume the independence between rare variants while others assume a mild correlation between rare variants in evaluating rare variants in association studies with common diseases.⁸³ Ignoring LD in rare variants may introduce bias or loss of power in association testing. In this study, we conducted simulation studies using independent rare variants and those simulated based on a genomic region on chromosome 21 in the human 1000 Genomes Project. We found that a large proportion (65%) of these simulated rare variants had no LD (r -squared < 0.01), while 32% displayed low pairwise LD (R^2 between 0.01 and 0.05), and about 4.5% showed R^2 in 0.05-0.33. By comparing the performance of the proposed method in clustering independent rare variants and rare variants with LD in simulation studies, we found that the proposed method performs better in clustering independent rare variants when the differences between the true means of clusters were relatively small, even if the multiple-variant model provided beta coefficients accounting for LD between rare variants. In addition, the multiple-variant model requires individual-level genotype data. Extending our algorithm to account for LD using existing summary data from large GWAS will be more cost-effective.

In summary, the proposed clustering algorithm identifies risk and/or protective rare variants of distinct magnitudes according to summary statistics of SNP-trait associations. The proposed method can be easily applied to summary statistics from emerging large-scale rare variants GWAS to identify and group trait-associated rare variants into null and signal groups of discrete effect magnitudes. Therefore, this proposed method may facilitate the identification of potentially causal rare variant clusters in genomic regions and ultimately help understand the genetic architecture underlying human complex traits for the discovery of drug targets and the design of gene therapy.

3.5 Tables

Table 3.1 Summary of six simulation scenarios with 1000 replicates

	Number of groups of rare variants	The ratio of true beta effects between each group	Proportions of the number of rare variants among each group
Scenario 1	3	0:1:2	$1/3:1/3:1/3$
Scenario 2	3	-1:0:1	$1/3:1/3:1/3$
Scenario 3	4	-1:0:1:2	$1/4:1/4:1/4:1/4$
Scenario 4	5	-2:-1:0:1:2	$1/5:1/5:1/5:1/5:1/5$
Scenario 5	2	1:2	$1/2:1/2$
Scenario 6	2	-1:1	$1/2:1/2$

Table 3.2 MSE of estimated clusters' means for a continuous trait with the absence of LD in simulation studies

Scenario	N=5000	N=15000	N=25000
Scenario 1	0.0102	0.00329	0.00169
Scenario 2	0.0102	0.00221	0.000868
Scenario 3	0.0124	0.00366	0.00179
Scenario 4	0.0127	0.00461	0.00228
Scenario 5	0.00535	0.0024	0.00141
Scenario 6	0.00574	0.000633	0.000143

Table 3.3 Summary of clustering results for rare variants within the combined signal regions of BP traits

Trait	# variants	# clusters	Mu	Phi	# variants of clusters	P-value ANOVA	P-value MWU
SBP	58	3	0/-0.0826/0.0573	0.719/0.178/0.102	46/8/4	4.26e-18	0.559
DBP	45	3	0/-0.0924/0.0471	0.718/0.165/0.117	37/5/3	1.68e-11	0.338
PP	55	5	0/-0.0438/- 0.0968/0.334/0.173	0.673/0.134/0.0654/0.0313/0.0966	43/4/2/1/5	1.39e-19	0.976
HTN	59	3	0/-2.785/4.646	0.802/0.147/0.0516	48/8/3	3.66e-16	0.32

Table 3.4 Summary of clustering results for rare variants within the signal genes of SBP

Gene	Chr	Start (bp)	End (bp)	# variants	# clusters	mu	phi	# variants of clusters	P-value _ANOVA	P-value MWU
DBH	9	136501569	136523555	27	2	0/-0.082	0.75/0.25	21/6	5.18e-08	0.785
NPR1	1	153652129	153665650	13	3	0/-0.0846/0.164	0.727/0.198/0.0753	10/2/1	6.02e-05	0.322
PLCB3	11	64021930	64034975	18	2	0/0.0529	0.716/0.284	15/3	2.33e-05	0.934

CHAPTER 4 ASSOCIATION ANALYSIS OF MITOCHONDRIAL HETEROPLASMIC VARIANTS AND CARDIOMETABOLIC TRAITS

4.1 Introduction

Mitochondria are the energy powerhouses of the cells. They are the center of energy metabolism and play critical roles in many cellular activities such as reactive oxygen species (ROS) production, Ca²⁺ levels, and apoptosis.^{3; 6; 92} Mitochondrial dysfunction has been implicated in the pathogenesis of multiple diseases including cardiovascular diseases.⁶ However, the underlying mechanisms between mitochondrial dysfunction and cardiovascular are complex. Recent advances in animal models indicated that mitochondrial dysfunction is likely to occur following the pathogenesis of atherosclerosis.⁹³

Mitochondria have their genome (mtDNA) which is maternally inherited and present in up to thousands of copies per cell. mtDNA encodes thirteen major genes for proteins of energy production in the oxidative phosphorylation pathway (OXPHOS). Several severe mitochondrial diseases are caused by maternally inherited, rare single nucleotide variants in mtDNA.⁹⁴ The rare mitochondrial disease often involves multiple organs of the human body including the cardiovascular system. Recent studies have shown that mtDNA inherited variants are also associated with common cardiometabolic diseases (CMDs) including hypertension and diabetes.^{30; 32; 95} Heteroplasmy refers to a

phenomenon in which a mixed population of mtDNA molecules (due to multiple alleles at a single site) are present in a cell. The role of heteroplasmy in cardiovascular disease is poorly understood owing to the lack of deep-sequencing of a large number of human genomes.

With the advent of next-generation sequencing technology, a large number of human genomes including mtDNA have been deep-sequenced, making it possible to investigate the role of heteroplasmy in human disease. Studies using family data showed that heteroplasmy is both inherited and arises somatically.^{52; 96} Multiple studies found that heteroplasmy is ubiquitous in the human population. However, 98% of heteroplasmic variants are rare and only present in one (i.e., singleton) or a few individuals^{24-26; 52; 96}, and in addition, their variant allele fractions (VAFs) are low in individuals. We recently proposed a comprehensive framework for association analyses of rare heteroplasmic variants identified in whole genome sequencing (WGS). In this study, we applied the proposed framework to six large cohorts with WGS data. We performed cohort- and ancestry-specific association analyses of heteroplasmic variants with four CMDs including obesity, hypertension, diabetes, and hyperlipidemia, and the continuous traits related to these CMDs.

4.2 Methods

4.2.1 Study participants

This study included 16,882 participants with WGS from six longitudinal cohort studies of multiple ancestries (Supplemental Table 1): the Atherosclerosis Risk in Communities study⁵⁴ (ARIC) (n=3,452), the Coronary Artery Risk Development in Young Adults Study (CARDIA)⁹⁷ (n=3,346), the Cardiovascular Health Study (CHS)⁵⁸ (n=3,341), the Framingham Heart Study (FHS)⁵⁵⁻⁵⁷ (n=1,633), the Jackson Heart Study (JHS)⁵⁹ (n=2,196), the Multi-Ethnic Study of Atherosclerosis Study (MESA)⁹⁸ (n=2,941). Because FHS and JHS consist of family data and our framework is proposed for uncorrelated individuals, we randomly select maternally uncorrelated individuals from these two cohorts. FHS only includes participants of European ancestry while JHS consists of only African Americans. Cohorts recruited mostly middle-aged participants (mean age ranging from 58 to 69) while CHS recruited older participants (mean age 74 years) at the baseline. We excluded two duplicated individuals from JHS and eight from MESA because they overlapped with ARIC.

4.2.2 Identification of heteroplasmy

Four TOPMed sequencing centers performed whole genome sequencing of these participating cohorts with an average coverage of 39-fold on the nuclear genome and ~3000 on the mtDNA.⁴¹ All participants for a given cohort were

sequenced at the same center. Data acquisition, DNA library construction, and data processing methods are described in detail elsewhere (<https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-methods-freeze-8>). The alignment of sequencing reads and handling of BAM files were described in detail in a previous study.⁵² The detailed quality control procedures of mtDNA sequencing were described previously⁵². We applied MToolBox⁴² to all participating cohorts (WGS TOPMed Freeze 8, released in February 2019, GRCH38)⁹⁹ with the reference mtDNA sequence, the revised Cambridge Reference Sequence (rCRS).¹⁶ Based on sequencing data from three FHS participants in a trio, we selected the 5%-95% of thresholds to identify heteroplasmy.⁵² That is, an mtDNA variant was considered a heteroplasmy if $5\% \leq \text{VAF} \leq 95\%$, and a variant is considered a homoplasmy if $\text{VAF} > 95\%$. Of note, these thresholds are narrower compared to what is used in Chapter 2 (3%-97%), because a recent study found narrower thresholds are likely to reduce the effect of NUMT on heteroplasmy identification.¹⁰⁰ Based on VAF and a pre-specified interval of 5%-95%, we consider three types of the genetic coding of heteroplasmic variants. For the i^{th} individual at the j^{th} site, the binary coding of a heteroplasmic variant is $G_{ij}^{(1)} = \begin{cases} 1 & \text{if } \text{VAF}_{ij} \in (0.05, 0.95) \\ 0 & \text{o. w.} \end{cases}$. The second coding is to incorporate variant allele fraction (VAF) as $G_{ij}^{(2)} = \begin{cases} \text{VAF}_{ij} & \text{if } \text{VAF}_{ij} \in (0.05 - 0.95) \\ 0 & \text{o. w.} \end{cases}$. Mito-score is a measure of local constraint, derived from the assessment of local intolerance to base/amino acid

substitutions. Each base is assigned a score between 0-1, and the higher the score the more locally constrained the base is. The mito-score coding is as $G_{ij}^{(3)} =$

$$\begin{cases} \text{MitoScore}_{ij} & \text{if } VAF_{ij} \in (0.05, 0.95) \\ 0 & \text{o.w.} \end{cases}$$

4.2.3 Cardiometabolic traits

We analyzed cross-sectional CMD traits, i.e., these traits were mapped to the health exams when blood was drawn for DNA extraction for mtDNA CN estimates. We focused on four continuous and four binary CMD phenotypes in primary analyses including body mass index (BMI), obesity, systolic blood pressure (SBP), hypertension, blood glucose (BG), diabetes, low-density lipoprotein (LDL), and hyperlipidemia. LDL (mg/dL) was calculated as $(TC - HDL - TRIG/5)$ in individuals with TRIG <400 mg/dL using imputed TC values.¹⁰¹ LDL values were log-transformed to approximate normality. The other continuous outcome variables were not transformed.

A therapeutic indication was provided for medication treatment in most, but not all, of the TOPMed cohorts. SBP and LDL variables are calibrated for medication treatment. For participants with hypertension treatment for lowering high blood pressure, we added 15 mmHg to the measured SBP. For participants without hypertension treatment, the measured SBP was used in the analysis. For

participants with lipid treatment, TC values were calculated as the measured TC divided by 0.8, and participants were removed if their TRIG values were larger than 400 mg/dl. Then LDL was calculated by the formula described in the above paragraph. For participants without lipid treatment, their LDL levels were calculated using their measured TC and TRIG levels. In the analysis of BG, we removed participants with measured BG levels ≥ 126 mg/dL or with diabetes treatment.

Obesity was defined as body mass index (BMI) ≥ 30 (kg/m²). Hypertension (HTN) was defined as systolic blood pressure (SBP) ≥ 140 mmHg, diastolic blood pressure (DBP) ≥ 90 mmHg, or the use of antihypertensive medication(s). Diabetes was defined as having a fasting blood glucose level of ≥ 126 mg/dL or currently receiving medications to lower blood glucose levels (BG) to treat diabetes. Hyperlipidemia was defined as fasting total cholesterol (TC) ≥ 200 mg/dL or triglyceride (TRIG) ≥ 150 mg/dL, or the use of any lipid-lowering medication.

4.2.4 Association analyses of rare heteroplasmic variants with cardiometabolic traits by gene-based tests

We only include rare heteroplasmic variants with population-level frequency $MAF_{Hj} = \frac{1}{n} \sum_i^n G_{ij}^{(1)} < 0.01$ in this study. We adopt four gene-based

tests for the association analyses: burden test, SKAT, SKAT-O, and ACAT-O.

The burden test collapses the heteroplasmic variants in the target region into a single burden statistic. SKAT is a variance component test that is more powerful than the burden test when the proportion of causal rare variants in the region is small. SKAT-O is an omnibus combining the burden test and SKAT to provide robust results with different proportions of causal variants and effect directions of variants in the target region. ACAT-O is a p-value combination method that is commonly used in genetic association studies. ACAT-O combines the p-values of a burden test and a SKAT by $Q_{ACAT-O} = \frac{1}{2} [\tan\{(0.5 - p_{burden})\} +$

$\tan\{(0.5 - p_{SKAT})\}]$. The p-value of ACAT-O is $p_{ACAT-O} \approx \frac{1}{2} - \frac{\arctan(Q_{ACAT-O})}{\pi}$. For

each of the tests, we consider equal weights of heteroplasmic variants within target regions. These four methods are described thoroughly in Chapter 2.

Covariates were incorporated by these methods. Age and sex are included as covariates in the model for all eight CMD traits. BMI is included as a covariate for the CMD traits except for BMI and obesity. Age squared is included for the traits of medication-adjusted SBP and hypertension. We adjust for visit year as a batch effect for FHS and JHS, because heteroplasmic burden is significantly associated with visit year in these two cohorts. Additionally, current smoking status is included as an adjustment variable except for ARIC and CARDIA due to unavailability. We run two sets of association analyses: 1. associations between the CMD outcomes and the heteroplasmic variants for the whole mtDNA genome; 2. associations between the CMD outcomes and the heteroplasmic

variants within the sixteen mtDNA genes/regions. We perform only the burden test for the first set of analyses, and we perform all four gene-based tests for the second set of analyses. We first perform cohort- and ancestry-specific analyses to obtain score statistics of heteroplasmic variants. Then for the two ancestries, we meta-analyze the results across the cohorts to obtain ancestry-specific results. We use a score statistics combination method to meta-analyze the results from ancestry-specific cohorts for the burden, SKAT, and SKAT-O methods implemented in the seqMeta R package. We adopt Bonferroni correction to control for sixteen mtDNA gene regions and set the significance level to be $\alpha = 0.05/16 \approx 0.00313$. We do not control for multiple CMD traits or multiple gene-based tests because they are highly correlated. It is too conservative to control traits and tests by Bonferroni correction. We consider that the heteroplasmic variants within a gene are associated with a CMD trait if at least one of the four tests provides significant results ($P \leq 0.05/16$) by one of the three coding definitions.

4.3 Results

4.3.1 Participants Characteristics

Most of the cohorts included mainly middle-aged participants except for CHS (mean age 74 years). These cohorts included more women (51%-64%) than men except for the FHS (47% women) due to the restriction to unrelated

samples according to maternal lineage information (Table 1). We observed heterogeneity in disease prevalence for the CMDs across the cohorts and two ancestries (i.e., AA versus EA). For example, the prevalence of obesity was higher in AA participants (51.3%) than EA participants in CARDIA (26.8%) (Chi-squared $P < 0.00001$). The prevalence of hypertension was 75.6% in EA participants in ARIC while it was 14.7% in EA participants in CARDIA (Chi-squared $P < 0.00001$), likely due to the large difference in age distributions (mean age is 58 in ARIC and 46 in CARDIA).

4.3.2 Heteroplasmy distribution

Due to the small sample size, we observed fewer heteroplasmic variants among AA participants than EA participants. For example, we observed 100 heteroplasmic variants within the *CYB* gene among the EA participants ($n=2,684$); while we observed only 30 heteroplasmic variants for the AA participants in CHS ($n=657$). The counts of heteroplasmic variants were also different across genes due to different gene lengths. For instance, we observed 121 heteroplasmic variants within the *ND5* gene (length of 1,812 base pairs, frequency= $121/1812 \approx 6.7\%$) among EA participants; while we observed only 16 within ATP8 gene (length of 207 base pairs, frequency= $16/207 \approx 7.7\%$) in CHS. But the differences in the frequency of heteroplasmic variants within these two genes were not statistically significant. (Chi-squared $P=0.67$).

4.3.3 Association between heteroplasmic burden and year of examination

We found that the year of blood draw of FHS and JHS individuals was significantly associated with the heteroplasmic burden (Table 4.3). Therefore, we additionally adjusted for the year of blood draw as a covariate for these two cohorts in gene-based tests.

4.3.4 Associations between heteroplasmy and CMD traits

We identified twelve pairs of gene-trait associations ($P \leq 0.05/16$). For all significant associations, the effect directions were consistent across the three coding definitions using the burden method (Table 3, Supplementary Tables 1-16). For example, for coding definition 1, the heteroplasmic burden within the mitochondrially encoded Cytochrome C Oxidase II (CO2) gene was associated with lower odds of obesity among AA participants (OR = 0.57, $P = 0.0015$) (Table 4.4). The burden test provided consistent effect direction by definition 2 (OR = 0.98, $P = 0.0016$) and definition 3 (OR = 0.15, $P = 0.0042$) as definition 1 (Table 4.4). The strongest association was obtained between hyperlipidemia and the heteroplasmic variants within the mitochondrially encoded Cytochrome C Oxidase I (CO1) gene among EA by burden test with coding definition 3 (OR = 0.28, $P = 3.4E-7$) (Table 4.4). Of note, the corresponding SKAT did not provide significant results after multiple testing corrections ($P = 0.02$) (Table 4.4). This may suggest this association was attributed to a large proportion of rare

heteroplasmic variants within the CO1 gene with small effect sizes and consistent effect direction. As expected, the heteroplasmic burden of the whole genome was also associated with lower odds of hyperlipidemia among EA participants (OR = 0.79, P = 0.00082, definition 3) (Table 4.4). However, we did not identify any hyperlipidemia-associated genes except for CO1, maybe because the hyperlipidemia-associated heteroplasmic were evenly distributed across these mtDNA genes and none of the genes contained enough trait-associated heteroplasmic variants to reach the significance level of 0.05/16 (Supplementary Tables 4.15-4.16). The twelve significant associations were detected by different gene-based tests. According to the simulation study of Chapter 2, this may imply inconsistent distributions and effect directions of CMD trait-associated heteroplasmic variants across different mtDNA gene-CMD trait associations. For example, the heteroplasmic variants within the CO1 gene were significantly associated with medication-adjusted LDL among EA using burden test (beta=-13.4, P=0.0006), while SKAT (P=0.068) did not provide significant results (Supplementary Table 4.14). We infer that a large proportion of heteroplasmic variants within the CO1 gene was associated with medication-adjusted LDL and most of the LDL-associated heteroplasmic variants had the same effect direction. The heteroplasmic variants within the CO3 gene were associated with medication-adjusted SBP among EA by SKAT (P=0.00051), while the burden test (P=0.4) did not show significant results (Supplementary Table 4.6). This may suggest that only a small proportion of heteroplasmic

variants within the CO3 gene were associated with medication-adjusted SBP, or around 50% of the SBP-associated heteroplasmic variants had opposite effect direction.

4.4 Discussion

In this study, we applied a framework that was developed in Chapter 2 to perform association analysis between heteroplasmic variants within sixteen mtDNA encoded genes/area and eight CMD traits in six TOPMed cohorts. We found the heteroplasmic within several mtDNA genes was associated with BMI, obesity, medication-adjusted SBP, restricted BG, diabetes, medication-adjusted LDL, and hyperlipidemia. In summary, were found fewer gene-trait associations among EA participants compared to AA participants, probably due to the smaller sample size (Table 1, Table 3). We found no overlapping heteroplasmy-associated CMD traits between AA participants and EA participants. This may partly be due to the different disease prevalence between AA and EA participants. In addition, the difference in sample size may also contribute to the observed difference.

Mitochondria are organelles of power production by generating ATP through oxidative phosphorylation. Mitochondria are also used for regulating cellular metabolism, cell death pathways, and calcium homeostasis. The

functions of mitochondria are strictly regulated by mtDNA. In this study, the strongest association was found between hyperlipidemia and heteroplasmic variants within the MT-CO1 gene. The MT-CO1 gene encodes the main subunit of the cytochrome c oxidase complex in complex IV which is the third and final enzyme of the electron transport chain of mitochondrial oxidative phosphorylation.¹⁰² Mutations in MT-CO1 have been associated with several diseases such as acquired idiopathic sideroblastic anemia, Complex IV deficiency, colorectal cancer, sensorineural deafness, and Leber's hereditary optic neuropathy.

To minimize false positives, we applied 5%-95% threshold on VAF to define heteroplasmic variants. Because our framework can easily incorporate weights of heteroplasmic variants at the individual level, we considered three coding definitions of heteroplasmy at the individual level in the association analyses. Definition 1 which was based on an indicator function was similar to the genetic coding of nDNA variants. Definition 2 incorporated the VAF and definition 3 incorporated the Mito-score which is a functional annotation score system of mtDNA. These three definitions provided comparable results, which made our findings more solid and reliable.

We identified both positive and negative associations between heteroplasmy and CMD traits (Table 4.4). This may imply that within the mtDNA genome, there exist heteroplasmic variants of different effect directions concerning CMD traits. Therefore, we plan to apply the clustering algorithm we developed in Chapter 3 to cluster heteroplasmic variants based on the variant-trait associations. There is increasing evidence that mtDNA and nDNA interact to influence disease traits. It is crucial to extend our study to account for the effects of nDNA by mtDNA-nDNA interaction analysis. In addition, the biological mechanisms of heteroplasmy on CMD traits remain to be studied. Leveraging mtDNA and other omics data such as gene expression data by causal mediation analysis may help unravel the potential causal pathways of heteroplasmy on CMD traits. In summary, our study provides valuable information to understand the effect of heteroplasmy on the pathological process of CMD traits, and further creates opportunities for therapy and drug development.

4.5 Tables

Table 4.1 Cohort-specific Characteristics for African Americans (A.) and European Americans (B.)

A.

Variable Mean (sd) or N (%)	ARIC (N=205)	CARDIA (N=1559)	CHS (N=657)	JHS (N=2196)	MESA (N=1089)
Ethnicity/Ancestry origin	AA	AA	AA	AA	AA
Age	58.7 (6.3)	44.4 (7.3)	73.7 (5.5)	56.5 (12.2)	60.9 (9.6)
Women	123 (60%)	922 (59.1%)	417 (63.5%)	1337 (60.9%)	577 (53%)
BMI	29.9 (7.1)	31.2 (7.4)	28.6 (5.5)	31.7 (7.2)	29.9 (5.4)
Obesity	89 (43.4%)	799 (51.3%)	226 (34.4%)	1169 (53.2%)	472 (43.3%)
Adj SBP	153.1 (25.8)	124 (19.4)	150.4 (23.6)	135.96 (19.5)	136.9 (23.6)
HTN	170 (82.9%)	543 (34.8%)	514 (78.2%)	1339 (61%)	724 (66.5%)
Adj FBG	100.4 (9.5)	93 (10.1)	97.3 (11.4)	90.6 (8.9)	90.1 (10.7)
DIAB	51 (24.9%)	159 (10.2%)	155 (23.6%)	470 (21.4%)	170 (15.6%)
Adj LDL	141.4 (51.9)	115.2 (36.8)	127.9 (41.4)	133.4 (40.2)	124.3 (37.2)
Hyperlipidemia	136 (66.3%)	680 (43.6%)	405 (61.6%)	1323 (60.3%)	592 (54.4%)
Smoking status	-	-	97 (14.8%)	1050 (47.8%)	197 (18.1%)

B.

Variable Mean (sd) or N (%)	ARIC (N=3247)	CARDIA (N=1787)	CHS (N=2684)	FHS (N=1633)	MESA (N=1825)
Ethnicity/Ancestry origin	EA	EA	EA	EA	EA
Age	58.2 (5.9)	45.5 (6.6)	74.2 (5.7)	59.4 (15.6)	61.6 (9.9)
Women	1651 (50.9%)	972 (54.4%)	1539 (57.3%)	771 (47.2%)	934 (51.2%)
BMI	27.5 (5)	27.7 (6.2)	26.4 (4.4)	27.7 (4.9)	27.8 (5)
Obesity	854 (26.3%)	479 (26.8%)	474 (17.7%)	449 (27.5%)	503 (27.6%)
Adj SBP	144.7 (21.5%)	113.7 (14.6)	142.8 (23.6)	132.3 (22.7)	127.3 (22.7)
HTN	2455 (75.6%)	262 (14.7%)	1783 (66.4%)	799 (48.9%)	903 (49.5%)
Adj FBG	101.2 (9.6)	92.8 (8.9)	98.6 (9.8)	97.3 (9.9)	87 (9.6)
DIAB	337 (10.4%)	76 (4.3%)	352 (13.1%)	181 (11.1%)	91 (5%)
Adj LDL	138.2 (40.6)	116.5 (32.7)	131.4 (39.3)	119.9 (33.5)	125.3 (31.5)
Hyperlipidemia	2278 (70.2%)	909 (50.9%)	1896 (70.6%)	1066 (65.3%)	1209 (66.3%)
Smoking status	-	-	273 (10.2%)	167 (10.2%)	194 (10.6%)

Table 4.2 Cohort-specific Distribution of Heteroplasmic variants Across Sixteen mtDNA genes for African Americans (A.) and European Americans (B.)

A.

	Gene_length	ARIC_AA	CARDIA_AA	CHS_AA	JHS	MESA_AA	Meta
D-loop	573	4	57	31	61	48	113
MT-RNR1	954	5	26	25	30	14	71
MT-RNR2	1558	6	39	31	55	30	123
MT-ND1	956	5	30	20	42	20	88
MT-ND2	1042	4	34	19	42	17	102
MT-CO1	1542	5	43	37	76	56	165
MT-CO2	684	0	19	13	30	15	66
MT-ATP8	207	0	6	6	14	5	26
MT-ATP6	635	0	30	14	46	34	94
MT-CO3	783	4	21	18	38	20	89
MT-ND3	346	0	12	7	10	7	30
MT-ND4L	297	1	10	4	8	5	21
MT-ND4	1371	2	34	24	49	21	113
MT-ND5	1812	13	47	38	79	47	166
MT-ND6	525	1	15	20	22	15	57
MT-CYB	1141	8	49	30	78	32	148
Total	14426	58	472	337	680	386	1472

B.

	Gene_length	ARIC_AA	CARDIA_AA	CHS_AA	JHS	MESA_AA	Meta
D-loop	573	80	44	80	50	54	149
MT-RNR1	954	55	19	75	38	35	169
MT-RNR2	1558	67	35	121	72	36	261
MT-ND1	956	43	26	53	29	29	130
MT-ND2	1042	32	25	67	36	33	155
MT-CO1	1542	97	41	94	65	55	260
MT-CO2	684	40	28	39	27	30	117
MT-ATP8	207	13	6	16	13	6	39
MT-ATP6	635	52	34	47	37	40	155
MT-CO3	783	39	28	59	25	32	138
MT-ND3	346	14	8	25	10	11	55
MT-ND4L	297	15	5	16	10	6	45
MT-ND4	1371	52	31	64	42	33	178
MT-ND5	1812	80	53	121	86	63	283
MT-ND6	525	32	29	43	26	17	101
MT-CYB	1141	83	58	100	56	62	254
Total	14426	794	470	1020	622	542	2489

Table 4.3 Associations of heteroplasmic burden with the year of blood draw

Cohort (N)	The P-value of heteroplasmic variants with the year of blood draw
African origin (N = 5,706)	
ARIC (N=205)	0.91
CARDIA (N=1559)	0.11
CHS (N=657)	0.64
JHS (N=2196)	0.013
MESA (N=1089)	0.79
European origin (N = 11,176)	
ARIC (N=3247)	0.08
CARDIA (N=1787)	0.68
CHS (N=2684)	0.44
FHS (N=1633)	0.0039
MESA (N=1825)	0.78

Table 4.4 Significant Associations between Heteroplasmic variants and CMD Traits across Sixteen mtDNA Genes by Definition 3 for continuous traits (A.) and binary traits (B.)

A.

Gene	ANC	Trait	Beta	SE	p_burden	p_skat	p_skato	p_acat
MT-CO3	EA	adjsbp	3.53	4.19	0.4	0.00051	0.00094	0.001
MT-CO1	EA	adjldl	-13.4	3.9	6.00E-04	0.068	0.0011	0.0012

B.

Gene	ANC	Trait	OR	95% CI	p_burden	p_skat	p_skato	p_acat
MT-CO1	AA	hyperlipid	0.28	(0.17, 0.46)	3.40E-07	0.02	0.018	6.90E-07
WG	AA	hyperlipid	0.79	(0.68, 0.9)	0.00082	-	-	-

Chapter 5 Summary and Future Work

5.1 Summary

Rare variants in nuclear DNA (nDNA) and mitochondrial DNA (mtDNA) are associated with disease traits. This dissertation aims to develop new strategies and apply the developed new strategies for association analyses of rare variants in both nuclear DNA and mitochondrial DNA.

In Chapter 1, we outlined the three projects in this dissertation. We reviewed the role of mitochondria in human life and mtDNA variants in human disease. Differently from nDNA variants, mtDNA variants display unique features, homoplasmy, and heteroplasmy. Most heteroplasmic variants are extremely rare. We also reviewed commonly used gene-based methods that have been developed in analyses of rare variants in nDNA. These gene-based methods remained to be evaluated for association analyses of mtDNA heteroplasmic variants, which motivated the development of Project 1 and Project 3 in this dissertation.

In Chapter 2, we proposed a framework for the association analysis of rare heteroplasmic variants. We considered pre-specified thresholds to identify heteroplasmy and two coding definitions to model heteroplasmy in association

analyses. We evaluated four commonly used gene-based tests and three burden-extension tests comprehensively through simulation studies. None of the tests is uniformly the most powerful across the simulation scenarios. The original burden test performs well when a large proportion of rare variants within the target region are associated with the trait and all of the trait-associated variants have the same effect direction. The burden-A method outperforms the original burden test when signal variants display opposite effect directions. The burden-S and burden-V methods improve the power further for the scenarios in that a large proportion of the variants are not trait-associated. Nevertheless, these three burden-extension methods are permutation-based and are more computationally intensive than the original burden test method. The SKAT outperforms the original burden test when only a small proportion of variants within the gene are trait-associated. The SKAT provides more robust results compared to the original burden test across different scenarios. The ACAT combines the p-values of the original burden test and SKAT efficiently. SKAT-O is underpowered based on the simulation compared to the ACAT method. In real data application, we found that heteroplasmic variants among several mtDNA genes, including RNR1, RNR2, CO1, CO2, and ND5, are associated with older age.

In summary, the burden-S has a larger power compared to the other methods in most of the simulation scenarios. Therefore, the burden-S method is recommended when the computational resource is sufficient. If the computational

resources are limited, the ACAT method is recommended, because ACAT is efficient and provides robust results.

In Chapter 3, we developed an efficient algorithm to cluster risk, protective, and/or neutral rare variants of signal gene regions that are identified by gene-based tests in association analyses of rare variants with traits. We cluster rare variants using their beta coefficients of variant-trait associations, e.g., in large GWAS. The clustering method is based on GMM with the EM algorithm. We proposed a maximum weighted sum of the Z-scores method to calculate the initial values of parameters. In the simulation study, we obtained the beta coefficients and their SE's using a multiple-variant model to account for potential LD structure. According to the simulation study, the proposed methods can efficiently cluster independent rare variants or those with moderate LD I when the study has a large sample size (e.g., in large GWAS and meta-analysis).

In Chapter 4, by using the framework developed in Chapter 2, we performed association analyses of heteroplasmic variants with CMD traits in six TOPMed cohorts, including ARIC, CARDIA, CHS, FHS, JHS, and MESA. The CMD traits included BMI, obesity, SBP, hypertension, blood glucose, diabetes, low-density lipoprotein, and hyperlipidemia. Heteroplasmic variants were identified from whole genome sequencing data generated by TOPMed

sequencing centers. We performed cohort- and ancestry-specific association analyses of heteroplasmic variants and CMD traits using the burden, SKAT, SKAT-O, and ACAT-O methods. We then conducted ancestry-specific meta-analyses. We also conducted meta-analyses across ancestries. We found heteroplasmic variants within several genes, including *CO1*, and *CO3*. are associated with obesity, diabetes, and/or hyperlipidemia.

5.2 Future Work

5.2.1 mtDNA Genotype Simulation

Genotype simulation plays an important role in genetic studies, such as evaluating the performance of novel statistical methods of genetics. To the best of our knowledge, unlike nDNA, there is no reliable software for mtDNA genotype simulation. To this end, we plan to develop software to simulate mtDNA VAF by mimicking the inheritance pattern of mtDNA using real mtDNA sequencing data as a reference. mtDNA displays unique features. First, mtDNA is maternally inherited; second, there are multiple mtDNA copies in a human cell; third, somatic mutation is occurred more frequently in mtDNA compared to nDNA, hence heteroplasmic burden is highly associated with age. Therefore, these three features are key points to be considered in software development.

5.2.2 Association Analysis of Heteroplasmy Using Correlated Data

In Chapter 2, we developed a framework for association analysis of heteroplasmy in independent individuals. However, many cohorts consist of correlated individuals of family members, such as FHS and JHS. Therefore, extending our framework to correlated data is one of the next steps. There are several potential challenges in this extension. First, because mtDNA is maternally inherited, the mitochondrial genetic correlation between maternal ancestors and offspring may cause computational issues in statistical modeling. Second, accounting for both nDNA and mtDNA genetic correlation may be computationally intensive.

5.2.3 Accounting for LD Structure Using Summary Statistics in Rare Variants Clustering

In Chapter 3, we developed an algorithm to cluster rare variants. To account for LD, we fitted a multiple-variant model to obtain beta coefficients and SE's in the simulation study. Fitting a multiple-variant model to account for an LD structure between rare variants requires individual-level genotype data, which may be costly to obtain genotypes and time-consuming to conduct gene-based tests. However, summary statistics are widely available from large GWAS studies. Therefore, accounting for LD using summary statistics from large GWAS is attractive and cost-effective. To the best of our knowledge, there is a method

developed recently for LD adjustment using summary data.¹⁰³ We plan to adopt this method and assess its performance in rare variants clustering by simulation studies.

APPENDIX A: SUPPLEMENTARY MATERIALS FOR CHAPTER 2

A.1 Supplementary Tables

Supplementary Table 2.1 Frequency of heteroplasmic sites in the CYB gene in simulation studies

Type	N	Singleton	Doubleton	3-5
All Heteroplasmy	116	97	17	7
Nonsynonymous heteroplasmy	66	55	9	4

We used the heteroplasmic sites in the mitochondrial Cytochrome b (MT-CYB) gene in European American participants (N=3,415) of the Atherosclerosis Risk in Communities (ARIC) Study for simulation studies. N, the total number of all or nonsynonymous heteroplasmic sites found; singleton, the heteroplasmic sites in single participants; doubleton, heteroplasmic sites in any two participants; 3-5, heteroplasmic sites in 3-5 participants. No heteroplasmic sites were found in more than five participants.

Supplementary Table 2.2 Gene-wise empirical type I error rates with 95% confidence interval for coding definition 2 using simulation data at $\alpha=0.001$

	Continuous Traits	Binary Traits (prevalence=20%)
Burden	0.86 (0.62, 1.16)	1.06 (0.79, 1.39)
Burden-A	1.14 (0.86, 1.48)	1.00 (0.74, 1.32)
Burden-S	1.34 (1.04, 1.7)	0.98 (0.73, 1.3)
Burden-V1	0.96 (0.71, 1.27)	0.00002 (0, 0.074)
Burden-V2	1.62 (1.23, 1.95)	0.00002 (0, 0.074)
SKAT	0.68 (0.47, 0.95)	0.06 (0.012, 0.18)
SKAT-O	0.80 (0.57, 1.09)	0.5 (0.32, 0.74)
ACAT	0.80 (0.57, 1.09)	0.72 (0.5, 1)

The type I error rate was represented as the ration of observed type I error to $\alpha=0.001$. Burden, the original burden test; Burden-A, adaptive burden test; Burden-S, the z-score weighting burden test; Burden-V1, variable threshold burden test with minimum p ; Burden-V2, variable threshold burden test with ACAT; SKAT, the sequence kernel association test; SKAT-O, the method combining the burden and SKAT; ACAT, the aggregated Cauchy association test combining the burden and SKAT. We simulated a continuous variable and a binary variable in response to heteroplasmic variants located in the mitochondrial cytochrome b (MT-CYB) gene in European American participants (N=3,415) of the Atherosclerosis Risk in Communities (ARIC) Study. We simulate 50,000 replicates for evaluating type I error.

Supplementary Table 2.3 Association analysis of heteroplasmic burden with the year of blood draw

Cohort (N)	Number of heteroplasmic sites	The P-value of heteroplasmic variants with the year of blood draw
African origin (n = 5456)		
ARIC (N=241)	162	0.63
CHS (N=705)	673	0.55
JHS (N=3404)	1590	1.77E-7
MESA (N=1106)	968	0.52
European origin (n = 12,051)		
ARIC (N=3415)	1501	0.32
CHS (N=2788)	1859	0.59
FHS (N=3992)	2158	7.01E-9
MESA (N=1856)	1236	0.66

Total n, the total number of heteroplasmic sites identified in a cohort. ARIC, Atherosclerosis Risk in Communities (ARIC) Study; FHS, Framingham Heart Study, CHS, Cardiovascular Health Study; JHS, Jackson Heart Study; MESA, Multi-Ethnic Study of Atherosclerosis. JHS includes participants of African Americans.

Supplementary Table 2.4 Distribution of heteroplasmic variants in the five cohorts

	ARIC_AA		ARIC_EA		CHS_AA		CHS_EA		FHS		JHS		MESA_AA		MESA_EA	
	N*	%*	n	%	n	%	n	%	n	%	n	%	n	%	n	%
<i>D-loop</i>	31	22.96	228	16.61	116	18.68	224	13.09	266	13.41	207	14.28	158	18.57	216	18.88
<i>RNR1</i>	8	5.93	87	6.34	35	5.64	129	7.54	146	7.36	59	4.07	37	4.35	56	4.90
<i>RNR2</i>	11	8.15	122	8.89	61	9.82	207	12.10	201	10.14	130	8.97	60	7.05	88	7.69
<i>ND1</i>	9	6.67	73	5.32	32	5.15	85	4.97	96	4.84	79	5.45	47	5.52	61	5.33
<i>MND2</i>	5	3.70	66	4.81	34	5.48	94	5.49	124	6.25	80	5.52	46	5.41	77	6.73
<i>CO1</i>	7	5.19	137	9.98	60	9.66	155	9.06	197	9.93	156	10.76	99	11.63	101	8.83
<i>CO2</i>	0	0.00	55	4.01	27	4.35	61	3.57	78	3.93	66	4.55	30	3.53	46	4.02
<i>ATP8</i>	0	0.00	18	1.31	8	1.29	25	1.46	34	1.71	25	1.72	17	2.00	10	0.87
<i>ATP6</i>	4	2.96	71	5.17	22	3.54	80	4.68	103	5.19	82	5.66	56	6.58	64	5.59
<i>CO3</i>	6	4.44	75	5.46	33	5.31	94	5.49	89	4.49	83	5.72	33	3.88	55	4.81
<i>ND3</i>	1	0.74	25	1.82	13	2.09	34	1.99	32	1.61	28	1.93	14	1.65	18	1.57
<i>ND4L</i>	1	0.74	24	1.75	6	0.97	26	1.52	28	1.41	18	1.24	10	1.18	14	1.22
<i>ND4</i>	9	6.67	89	6.48	29	4.67	103	6.02	124	6.25	90	6.21	51	5.99	74	6.47
<i>ND5</i>	25	18.52	140	10.20	66	10.63	190	11.10	238	12.00	170	11.72	98	11.52	122	10.66
<i>ND6</i>	2	1.48	47	3.42	28	4.51	61	3.57	67	3.38	47	3.24	24	2.82	40	3.50
<i>CYB</i>	16	11.85	116	8.45	51	8.21	143	8.36	160	8.07	130	8.97	71	8.34	102	8.92
Total	135	100	1373	100	621	100	1711	100	1983	100	1450	100	851	100	1144	100

*, number (n) of heteroplasmic variants in each gene/area; the proportion of heteroplasmic variants (%) of all variants in an ancestry-specific cohort. ARIC, Atherosclerosis Risk in Communities (ARIC) Study; FHS, Framingham Heart Study, CHS, Cardiovascular Health Study; JHS, Jackson Heart Study; MESA, Multi-Ethnic Study of Atherosclerosis. JHS includes participants of African Americans.

Supplementary Table 2.5 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 1 from Fisher's method meta-analysis for participants of African American ancestry

mtDNA region	P-values							
	Burden	Burden-A	Burden-S	Burden-V1	Burden-V2	SKAT	SKAT-O	ACAT
<i>D-loop</i>	0.21	0.12	0.19	0.85	0.84	0.27	0.31	0.26
<i>MT-RNR1</i>	0.21	0.038	0.1	0.1	0.11	0.22	0.29	0.22
<i>MT-RNR2</i>	0.006	0.001	0.0026	0.001	0.0013	0.18	0.022	0.026
<i>MT-ND1</i>	0.46	0.2	0.18	0.033	0.028	0.32	0.25	0.38
<i>MT-ND2</i>	0.13	0.15	0.15	0.1	0.091	0.22	0.18	0.13
<i>MT-CO1</i>	0.22	0.18	0.33	0.18	0.19	0.34	0.18	0.21
<i>MT-CO2</i>	0.35	0.45	0.44	0.79	0.79	0.69	0.41	0.69
<i>MT-ATP8</i>	0.56	0.46	0.39	0.19	0.21	0.22	0.43	0.35
<i>MT-ATP6</i>	0.78	0.54	0.6	0.58	0.51	0.53	0.65	0.74
<i>MT-CO3</i>	0.2	0.16	0.34	0.96	0.95	0.91	0.34	0.41
<i>MT-ND3</i>	0.58	0.43	0.81	0.44	0.43	0.53	0.71	0.55
<i>MT-ND4L</i>	0.81	0.71	0.39	0.37	0.38	0.38	0.55	0.6
<i>MT-ND4</i>	0.33	0.47	0.48	0.97	0.96	0.82	0.45	0.57
<i>MT-ND5</i>	0.16	0.17	0.24	0.29	0.31	0.64	0.32	0.32
<i>MT-ND6</i>	0.011	0.04	0.04	0.051	0.054	0.065	0.023	0.019
<i>MT-CYB</i>	0.062	0.047	0.15	0.53	0.62	0.57	0.17	0.15

We perform cohort-specific association analyses between heteroplasmic mutations and age. Meta-analysis was performed with Fisher's method to combine the p-values. Burden, the original burden test; Burden-A, adaptive burden test; Burden-S, the z-score weighting burden test; Burden-V1, variable threshold burden test with minimum p; Burden-V2, variable threshold burden test with ACAT; SKAT, the sequence kernel association test; SKAT-O, the method combining the burden and SKAT; ACAT, the aggregated Cauchy association test combining the burden and SKAT. MT-RNR1/RNR2, the two ribosomal RNA genes in mitochondrial DNA; MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; MT-CO1/CO2/CO3, the mitochondrial cytochrome c oxidase I, II, and III genes; MT-CYB, the mitochondrial cytochrome b gene; MT-APT6/ATP8, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.6 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 2 from Fisher's method meta-analysis for participants of African American ancestry

mtDNA region	P-values							
	Burden	Burden-A	Burden-S	Burden-V1	Burden-V2	SKAT	SKAT-O	ACAT
<i>D-loop</i>	0.34	0.26	0.39	0.97	0.93	0.93	0.57	0.51
<i>MT-RNR1</i>	0.086	0.047	0.22	0.14	0.15	0.49	0.036	0.15
<i>MT-RNR2</i>	0.0019	0.00013	0.00042	0.00054	4.00E-04	0.033	0.28	0.0061
<i>MT-ND1</i>	0.22	0.17	0.11	0.076	0.062	0.32	0.26	0.16
<i>MT-ND2</i>	0.2	0.24	0.16	0.23	0.19	0.46	0.26	0.25
<i>MT-CO1</i>	0.086	0.093	0.18	0.45	0.53	0.44	0.41	0.19
<i>MT-CO2</i>	0.34	0.39	0.31	0.72	0.71	0.62	0.85	0.56
<i>MT-ATP8</i>	0.87	0.6	0.56	0.3	0.32	0.51	0.2	0.79
<i>MT-ATP6</i>	0.74	0.51	0.49	0.63	0.56	0.6	0.42	0.76
<i>MT-CO3</i>	0.42	0.27	0.42	0.75	0.78	0.88	0.27	0.68
<i>MT-ND3</i>	0.77	0.5	0.78	0.43	0.43	0.37	0.24	0.59
<i>MT-ND4L</i>	0.84	0.74	0.39	0.46	0.5	0.41	0.58	0.64
<i>MT-ND4</i>	0.18	0.39	0.36	0.94	0.94	0.78	0.41	0.32
<i>MT-ND5</i>	0.08	0.17	0.21	0.053	0.23	0.66	0.53	0.16
<i>MT-ND6</i>	0.023	0.052	0.051	0.099	0.1	0.24	0.12	0.057
<i>MT-CYB</i>	0.032	0.043	0.16	0.49	0.41	0.73	0.91	0.1

We perform cohort-specific association analyses between heteroplasmic mutations and age. Meta-analysis was performed with Fisher's method to combine the p-values. Burden, the original burden test; Burden-A, adaptive burden test; Burden-S, the z-score weighting burden test; Burden-V1, variable threshold burden test with minimum p; Burden-V2, variable threshold burden test with ACAT; SKAT, the sequence kernel association test; SKAT-O, the method combining the burden and SKAT; ACAT, the aggregated Cauchy association test combining the burden and SKAT. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.7 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 1 from the fixed-effect meta-analysis for participants of African American ancestry

Gene/region	BETA	SE	95% LCL	95% UCL	P
<i>D-loop</i>	0.23	0.11	0.016	0.45	0.036
<i>MT-RNR1</i>	0.6	0.39	-0.17	1.37	0.13
<i>MT-RNR2</i>	1.05	0.31	0.45	1.65	0.00059
<i>MT-ND1</i>	0.49	0.37	-0.23	1.21	0.18
<i>MT-ND2</i>	0.93	0.46	0.03	1.84	0.043
<i>MT-CO1</i>	0.33	0.28	-0.21	0.87	0.23
<i>MT-CO2</i>	0.85	0.54	-0.2	1.9	0.11
<i>MT-ATP8</i>	0.13	0.95	-1.72	1.99	0.89
<i>MT-ATP6</i>	0.21	0.48	-0.74	1.16	0.67
<i>MT-CO3</i>	0.55	0.48	-0.39	1.49	0.25
<i>MT-ND3</i>	1.06	0.77	-0.45	2.58	0.17
<i>MT-ND4L</i>	0.7	1.02	-1.3	2.7	0.49
<i>MT-ND4</i>	0.55	0.34	-0.12	1.23	0.11
<i>MT-ND5</i>	0.44	0.2	0.047	0.84	0.028
<i>MT-ND6</i>	1	0.53	-0.041	2.04	0.06
<i>MT-CYB</i>	0.93	0.35	0.24	1.61	0.0079

We perform cohort-specific association analyses between heteroplasmic mutations and age. Meta-analysis was performed with the fixed-effects inverse variance method. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.8 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 2 from the fixed-effect meta-analysis for participants of African American ancestry

Gene/region	BETA	SE	95% LCL	95% UCL	P
<i>D-loop</i>	0.0095	0.0056	-0.0015	0.021	0.091
<i>MT-RNR1</i>	0.032	0.015	0.0017	0.062	0.039
<i>MT-RNR2</i>	0.039	0.011	0.018	0.06	0.00025
<i>MT-ND1</i>	0.0082	0.013	-0.017	0.033	0.52
<i>MT-ND2</i>	0.021	0.014	-0.0076	0.049	0.15
<i>MT-CO1</i>	0.02	0.0098	0.00032	0.039	0.046
<i>MT-CO2</i>	0.031	0.018	-0.0043	0.065	0.086
<i>MT-ATP8</i>	-0.0036	0.029	-0.061	0.053	0.9
<i>MT-ATP6</i>	0.0094	0.016	-0.022	0.041	0.55
<i>MT-CO3</i>	0.021	0.015	-0.0091	0.05	0.17
<i>MT-ND3</i>	0.02	0.026	-0.031	0.072	0.43
<i>MT-ND4L</i>	0.019	0.037	-0.053	0.091	0.61
<i>MT-ND4</i>	0.028	0.013	0.0024	0.054	0.032
<i>MT-ND5</i>	0.022	0.008	0.006	0.037	0.0067
<i>MT-ND6</i>	0.043	0.019	0.0056	0.079	0.024
<i>MT-CYB</i>	0.031	0.011	0.0086	0.053	0.0066

We perform cohort-specific association analyses between heteroplasmic mutations and age. Meta-analysis was performed with the fixed-effects inverse variance method. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.9 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 1 from Fisher's method meta-analysis for participants of European American ancestry

mtDNA region	P-values							
	Burden	Burden-A	Burden-S	Burden-V1	Burden-V2	SKAT	SKAT-O	ACAT
<i>D-loop</i>	0.0091	0.018	0.022	0.14	0.58	0.02	0.004	0.011
<i>MT-RNR1</i>	2.40E-09	1.50E-08	2.00E-08	9.70E-05	0.0039	0.0041	1.60E-05	7.10E-09
<i>MT-RNR2</i>	1.60E-09	8.30E-09	3.80E-08	0.045	2.20E-06	0.0051	4.30E-06	1.80E-08
<i>MT-ND1</i>	0.14	0.06	0.037	0.49	0.37	0.21	0.085	0.11
<i>MT-ND2</i>	0.14	0.2	0.23	0.97	0.95	0.61	0.33	0.38
<i>MT-CO1</i>	2.50E-05	5.50E-06	5.30E-06	0.063	3.00E-04	0.039	0.00018	0.00014
<i>MT-CO2</i>	0.02	0.0059	0.01	0.0094	0.0075	0.0084	0.0018	0.0021
<i>MT-ATP8</i>	0.0021	0.0016	0.0025	0.1	0.1	0.051	0.0071	0.0036
<i>MT-ATP6</i>	0.47	0.17	0.16	0.81	0.82	0.83	0.71	0.61
<i>MT-CO3</i>	0.0032	0.0021	0.0015	0.044	0.11	0.029	0.012	0.0057
<i>MT-ND3</i>	0.29	0.0054	0.0042	0.81	0.76	0.099	0.086	0.087
<i>MT-ND4L</i>	0.77	0.73	0.66	0.74	0.73	0.35	0.6	0.47
<i>MT-ND4</i>	0.062	0.025	0.012	0.045	0.056	0.3	0.15	0.11
<i>MT-ND5</i>	0.0069	0.048	0.043	0.1	0.67	0.76	0.022	0.026
<i>MT-ND6</i>	0.83	0.4	0.34	0.26	0.27	0.15	0.32	0.39
<i>MT-CYB</i>	0.22	0.0045	0.0024	0.53	0.56	0.27	0.22	0.39

We perform cohort-specific association analyses between heteroplasmic mutations and age. Meta-analysis was performed with Fisher's method to combine the p-values. Burden, the original burden test; Burden-A, adaptive burden test; Burden-S, the z-score weighting burden test; Burden-V1, variable threshold burden test with minimum p; Burden-V2, variable threshold burden test with ACAT; SKAT, the sequence kernel association test; SKAT-O, the method combining the burden and SKAT; ACAT, the aggregated Cauchy association test combining the burden and SKAT. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.10 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 2 from Fisher's method meta-analysis for participants of European American ancestry

mtDNA region	P-values							
	Burden	Burden-A	Burden-S	Burden-V1	Burden-V2	SKAT	SKAT-O	ACAT
<i>D-loop</i>	0.027	0.016	0.011	0.29	0.83	0.18	0.01	0.07
<i>MT-RNR1</i>	2.10E-12	1.10E-08	1.10E-08	2.10E-05	5.90E-07	0.00029	0.00011	1.50E-11
<i>MT-RNR2</i>	1.20E-10	6.40E-08	2.10E-07	0.064	5.00E-06	0.058	0.00028	1.50E-09
<i>MT-ND1</i>	0.28	0.15	0.084	0.73	0.6	0.46	0.27	0.35
<i>MT-ND2</i>	0.093	0.18	0.19	0.96	0.94	0.88	0.055	0.33
<i>MT-CO1</i>	1.30E-06	2.10E-06	1.70E-06	0.019	8.20E-06	0.02	0.00065	7.50E-06
<i>MT-CO2</i>	0.0044	0.0096	0.01	0.013	0.0099	0.0012	6.70E-05	0.00025
<i>MT-ATP8</i>	0.0019	0.0058	0.011	0.13	0.13	0.15	0.15	0.0058
<i>MT-ATP6</i>	0.47	0.087	0.063	0.56	0.56	0.5	0.42	0.44
<i>MT-CO3</i>	0.022	0.018	0.014	0.27	0.25	0.28	0.19	0.15
<i>MT-ND3</i>	0.32	0.022	0.013	0.63	0.62	0.42	0.22	0.33
<i>MT-ND4L</i>	0.94	0.96	0.97	0.87	0.89	0.89	0.96	0.93
<i>MT-ND4</i>	0.003	0.0015	0.00036	0.018	0.019	0.056	0.15	0.0064
<i>MT-ND5</i>	0.0081	0.024	0.012	0.26	0.34	0.73	0.1	0.032
<i>MT-ND6</i>	0.59	0.08	0.042	0.16	0.16	0.078	0.37	0.19
<i>MT-CYB</i>	0.014	0.014	0.0044	0.39	0.72	0.097	0.64	0.037

We perform cohort-specific association analyses between heteroplasmic mutations and age. Meta-analysis was performed with Fisher's method to combine the p-values. Burden, the original burden test; Burden-A, adaptive burden test; Burden-S, the z-score weighting burden test; Burden-V1, variable threshold burden test with minimum p; Burden-V2, variable threshold burden test with ACAT; SKAT, the sequence kernel association test; SKAT-O, the method combining the burden and SKAT; ACAT, the aggregated Cauchy association test combining the burden and SKAT. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.11 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 1 from the fixed-effect meta-analysis for participants of European American ancestry

Gene/region	BETA	SE	95% LCL	95% UCL	P
<i>D-loop</i>	0.22	0.068	0.089	0.35	0.0011
<i>MT-RNR1</i>	1.32	0.26	0.8	1.83	5.00E-07
<i>MT-RNR2</i>	1.34	0.21	0.93	1.76	3.00E-10
<i>MT-ND1</i>	0.78	0.32	0.14	1.42	0.016
<i>MT-ND2</i>	0.68	0.3	0.086	1.27	0.025
<i>MT-CO1</i>	1.1	0.23	0.64	1.56	2.20E-06
<i>MT-CO2</i>	0.9	0.39	0.14	1.66	0.02
<i>MT-ATP8</i>	1.67	0.69	0.32	3.02	0.016
<i>MT-ATP6</i>	0.43	0.34	-0.24	1.11	0.21
<i>MT-CO3</i>	1.26	0.33	0.61	1.9	0.00015
<i>MT-ND3</i>	0.61	0.53	-0.44	1.65	0.26
<i>MT-ND4L</i>	-0.15	0.6	-1.33	1.03	0.8
<i>MT-ND4</i>	0.56	0.23	0.11	1.02	0.015
<i>MT-ND5</i>	0.45	0.17	0.13	0.77	0.0063
<i>MT-ND6</i>	0.26	0.4	-0.52	1.05	0.51
<i>MT-CYB</i>	0.43	0.21	0.012	0.84	0.044

We perform cohort-specific association analyses between heteroplasmic mutations and age. Meta-analysis was performed with the fixed-effects inverse variance method. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.12 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 2 from the fixed-effect meta-analysis for participants of European American ancestry

Gene/region	BETA	SE	95% LCL	95% UCL	P
<i>D-loop</i>	0.0083	0.0028	0.0028	0.014	0.0033
<i>MT-RNR1</i>	0.037	0.0059	0.025	0.048	4.80E-10
<i>MT-RNR2</i>	0.03	0.0048	0.021	0.04	2.90E-10
<i>MT-ND1</i>	0.014	0.0069	-4.30E-05	0.027	0.051
<i>MT-ND2</i>	0.015	0.0066	0.0022	0.028	0.022
<i>MT-CO1</i>	0.025	0.0051	0.015	0.035	9.20E-07
<i>MT-CO2</i>	0.022	0.0082	0.0062	0.038	0.0068
<i>MT-ATP8</i>	0.031	0.014	0.0042	0.057	0.023
<i>MT-ATP6</i>	0.0087	0.0073	-0.0056	0.023	0.23
<i>MT-CO3</i>	0.021	0.0069	0.0069	0.034	0.0031
<i>MT-ND3</i>	0.019	0.012	-0.0046	0.043	0.11
<i>MT-ND4L</i>	-0.0068	0.012	-0.031	0.017	0.58
<i>MT-ND4</i>	0.02	0.0057	0.0088	0.031	0.00047
<i>MT-ND5</i>	0.0098	0.0041	0.0018	0.018	0.017
<i>MT-ND6</i>	0.0067	0.0087	-0.01	0.024	0.44
<i>MT-CYB</i>	0.016	0.0052	0.0062	0.027	0.0016

We perform cohort-specific association analyses between heteroplasmic mutations and age. Meta-analysis was performed with the fixed-effects inverse variance method. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.13 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 1 from Fisher's method meta-analysis for all participants

mtDNA region	P-values							
	Burden	Burden-A	Burden-S	Burden-V1	Burden-V2	SKAT	SKAT-O	ACAT
<i>D-loop</i>	0.014	0.015	0.027	0.37	0.84	0.034	0.0095	0.02
<i>MT-RNR1</i>	1.10E-08	1.30E-08	4.20E-08	0.00012	0.0038	0.0072	6.20E-05	3.30E-08
<i>MT-RNR2</i>	2.50E-10	2.20E-10	2.40E-09	5.00E-04	5.90E-08	0.0073	1.60E-06	1.10E-08
<i>MT-ND1</i>	0.24	0.065	0.04	0.083	0.058	0.25	0.1	0.17
<i>MT-ND2</i>	0.091	0.14	0.15	0.32	0.3	0.4	0.23	0.2
<i>MT-CO1</i>	7.20E-05	1.50E-05	2.50E-05	0.062	0.00061	0.071	0.00037	0.00034
<i>MT-CO2</i>	0.042	0.018	0.028	0.044	0.036	0.036	0.0061	0.011
<i>MT-ATP8</i>	0.0091	0.006	0.0077	0.094	0.1	0.062	0.021	0.0097
<i>MT-ATP6</i>	0.73	0.31	0.32	0.82	0.78	0.8	0.82	0.81
<i>MT-CO3</i>	0.0053	0.003	0.0044	0.18	0.34	0.12	0.027	0.016
<i>MT-ND3</i>	0.47	0.016	0.023	0.72	0.69	0.21	0.23	0.19
<i>MT-ND4L</i>	0.92	0.86	0.61	0.63	0.63	0.4	0.7	0.64
<i>MT-ND4</i>	0.1	0.064	0.035	0.18	0.21	0.59	0.25	0.24
<i>MT-ND5</i>	0.0086	0.047	0.058	0.13	0.53	0.84	0.042	0.048
<i>MT-ND6</i>	0.052	0.082	0.072	0.071	0.076	0.055	0.044	0.044
<i>MT-CYB</i>	0.072	0.002	0.0032	0.64	0.71	0.44	0.16	0.22

We perform cohort-specific association analyses between heteroplasmic mutations and age. Meta-analysis was performed with Fisher's method to combine the p-values in all participants. Burden, the original burden test; Burden-A, adaptive burden test; Burden-S, the z-score weighting burden test; Burden-V1, variable threshold burden test with minimum p; Burden-V2, variable threshold burden test with ACAT; SKAT, the sequence kernel association test; SKAT-O, the method combining the burden and SKAT; ACAT, the aggregated Cauchy association test combining the burden and SKAT. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.14 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 2 from Fisher's method meta-analysis for all participants

mtDNA region	P-values							
	Burden	Burden-A	Burden-S	Burden-V1	Burden-V2	SKAT	SKAT-O	ACAT
<i>D-loop</i>	0.052	0.027	0.028	0.64	0.97	0.47	0.035	0.15
<i>MT-RNR1</i>	5.50E-12	1.20E-08	5.00E-08	4.00E-05	1.50E-06	0.0014	5.30E-05	6.30E-11
<i>MT-RNR2</i>	6.90E-12	2.20E-10	2.10E-09	0.00039	4.20E-08	0.014	0.00082	2.40E-10
<i>MT-ND1</i>	0.23	0.12	0.053	0.22	0.16	0.43	0.26	0.22
<i>MT-ND2</i>	0.093	0.18	0.14	0.55	0.49	0.77	0.075	0.29
<i>MT-CO1</i>	1.90E-06	3.20E-06	4.90E-06	0.049	5.80E-05	0.05	0.0025	2.10E-05
<i>MT-CO2</i>	0.011	0.025	0.021	0.053	0.042	0.0061	0.00061	0.0014
<i>MT-ATP8</i>	0.012	0.023	0.038	0.17	0.17	0.27	0.14	0.029
<i>MT-ATP6</i>	0.72	0.18	0.14	0.72	0.68	0.66	0.48	0.7
<i>MT-CO3</i>	0.053	0.031	0.036	0.53	0.51	0.59	0.2	0.33
<i>MT-ND3</i>	0.59	0.061	0.057	0.62	0.62	0.44	0.21	0.51
<i>MT-ND4L</i>	0.98	0.95	0.75	0.77	0.81	0.73	0.88	0.9
<i>MT-ND4</i>	0.0046	0.0049	0.0013	0.086	0.09	0.18	0.23	0.015
<i>MT-ND5</i>	0.0054	0.027	0.018	0.073	0.28	0.83	0.21	0.032
<i>MT-ND6</i>	0.072	0.027	0.015	0.082	0.082	0.093	0.18	0.06
<i>MT-CYB</i>	0.0039	0.0051	0.0058	0.51	0.66	0.26	0.9	0.024

We perform cohort-specific association analyses between heteroplasmic mutations and age. Meta-analysis was performed with Fisher's method to combine the p-values. Burden, the original burden test; Burden-A, adaptive burden test; Burden-S, the z-score weighting burden test; Burden-V1, variable threshold burden test with minimum p; Burden-V2, variable threshold burden test with ACAT; SKAT, the sequence kernel association test; SKAT-O, the method combining the burden and SKAT; ACAT, the aggregated Cauchy association test combining the burden and SKAT. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.15 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 1 from the fixed-effect meta-analysis of all participants

Gene/region	BETA	SE	95% LCL	95% UCL	P
<i>D-loop</i>	0.22	0.058	0.11	0.34	0.00012
<i>MT-RNR1</i>	1.1	0.22	0.67	1.52	3.80E-07
<i>MT-RNR2</i>	1.25	0.17	0.91	1.59	6.80E-13
<i>MT-ND1</i>	0.66	0.24	0.18	1.13	0.0067
<i>MT-ND2</i>	0.75	0.25	0.26	1.25	0.0027
<i>MT-CO1</i>	0.79	0.18	0.44	1.14	8.80E-06
<i>MT-CO2</i>	0.88	0.32	0.26	1.5	0.0052
<i>MT-ATP8</i>	1.14	0.56	0.044	2.23	0.041
<i>MT-ATP6</i>	0.36	0.28	-0.19	0.9	0.2
<i>MT-CO3</i>	1.03	0.27	0.5	1.57	0.00015
<i>MT-ND3</i>	0.75	0.44	-0.1	1.61	0.084
<i>MT-ND4L</i>	0.069	0.52	-0.95	1.08	0.89
<i>MT-ND4</i>	0.56	0.19	0.18	0.93	0.0035
<i>MT-ND5</i>	0.45	0.13	0.19	0.7	0.00058
<i>MT-ND6</i>	0.53	0.32	-0.097	1.15	0.098
<i>MT-CYB</i>	0.56	0.18	0.21	0.92	0.0018

We perform cohort-specific association analyses between heteroplasmic mutations and age. Meta-analysis was performed with the fixed-effects inverse variance method. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.16 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and age for coding definition 2 from the fixed-effect meta-analysis of all participants

Gene/region	BETA	SE	95% LCL	95% UCL	P
<i>D-loop</i>	0.0085	0.0025	0.0036	0.013	0.00065
<i>MT-RNR1</i>	0.036	0.0055	0.026	0.047	3.70E-11
<i>MT-RNR2</i>	0.031	0.0044	0.023	0.04	8.90E-13
<i>MT-ND1</i>	0.013	0.0061	0.00078	0.025	0.037
<i>MT-ND2</i>	0.016	0.006	0.0044	0.028	0.007
<i>MT-CO1</i>	0.024	0.0045	0.015	0.033	1.20E-07
<i>MT-CO2</i>	0.024	0.0075	0.0089	0.038	0.0016
<i>MT-ATP8</i>	0.024	0.013	-0.00025	0.049	0.052
<i>MT-ATP6</i>	0.0088	0.0066	-0.0042	0.022	0.18
<i>MT-CO3</i>	0.021	0.0063	0.0087	0.033	0.00081
<i>MT-ND3</i>	0.019	0.011	-0.0022	0.041	0.078
<i>MT-ND4L</i>	-0.0043	0.011	-0.027	0.018	0.7
<i>MT-ND4</i>	0.021	0.0052	0.011	0.032	4.50E-05
<i>MT-ND5</i>	0.012	0.0036	0.0052	0.019	0.00072
<i>MT-ND6</i>	0.013	0.0079	-0.0025	0.028	0.1
<i>MT-CYB</i>	0.019	0.0047	0.0095	0.028	6.70E-05

We perform cohort-specific association analyses between heteroplasmic mutations and age. Meta-analysis was performed with the fixed-effects inverse variance method. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.17 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 1 from Fisher's method meta-analysis for participants of African American ancestry

mtDNA region	P-values							
	Burden	Burden-A	Burden-S	Burden-V1	Burden-V2	SKAT	SKAT-O	ACAT
<i>D-loop</i>	0.28	0.22	0.18	0.36	0.37	0.19	0.22	0.26
<i>MT-RNR1</i>	0.4	0.37	0.067	0.54	0.61	0.53	0.49	0.49
<i>MT-RNR2</i>	0.57	0.57	0.28	0.35	0.39	0.2	0.37	0.37
<i>MT-ND1</i>	0.75	0.76	0.2	0.18	0.21	0.71	0.78	0.79
<i>MT-ND2</i>	0.31	0.38	0.21	0.14	0.17	0.69	0.48	0.42
<i>MT-CO1</i>	0.26	0.24	0.27	0.047	0.24	0.051	0.076	0.088
<i>MT-CO2</i>	0.011	0.01	0.0056	0.28	0.28	0.31	0.025	0.028
<i>MT-ATP8</i>	0.71	0.78	0.21	0.87	0.81	0.76	0.84	0.76
<i>MT-ATP6</i>	0.051	0.049	0.11	0.36	0.4	0.44	0.075	0.07
<i>MT-CO3</i>	0.86	0.88	0.19	0.86	0.89	0.13	0.34	0.3
<i>MT-ND3</i>	0.089	0.083	0.045	0.26	0.25	0.25	0.12	0.17
<i>MT-ND4L</i>	0.42	0.38	0.31	0.41	0.4	0.22	0.29	0.29
<i>MT-ND4</i>	0.1	0.12	0.029	0.11	0.14	0.22	0.18	0.15
<i>MT-ND5</i>	0.002	0.0011	0.0026	0.082	0.08	0.017	0.012	0.0038
<i>MT-ND6</i>	0.93	0.94	0.78	0.54	0.52	0.4	0.7	0.79
<i>MT-CYB</i>	0.44	0.46	0.21	0.19	0.19	0.34	0.56	0.44

We perform cohort-specific association analyses between heteroplasmic mutations and sex. Meta-analysis was performed with Fisher's method to combine the p-values. Burden, the original burden test; Burden-A, adaptive burden test; Burden-S, the z-score weighting burden test; Burden-V1, variable threshold burden test with minimum p; Burden-V2, variable threshold burden test with ACAT; SKAT, the sequence kernel association test; SKAT-O, the method combining the burden and SKAT; ACAT, the aggregated Cauchy association test combining the burden and SKAT. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.18 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 2 from Fisher's method meta-analysis for participants of African American ancestry

mtDNA region	P-values							
	Burden	Burden-A	Burden-S	Burden-V1	Burden-V2	SKAT	SKAT-O	ACAT
<i>D-loop</i>	0.071	0.055	0.044	0.54	0.51	0.25	0.095	0.095
<i>MT-RNR1</i>	0.27	0.26	0.12	0.44	0.45	0.51	0.088	0.43
<i>MT-RNR2</i>	0.26	0.27	0.2	0.17	0.21	0.18	0.13	0.24
<i>MT-ND1</i>	0.76	0.78	0.34	0.15	0.17	0.6	0.57	0.72
<i>MT-ND2</i>	0.28	0.34	0.21	0.093	0.12	0.64	0.59	0.43
<i>MT-CO1</i>	0.81	0.83	0.57	0.15	0.26	0.46	0.86	0.72
<i>MT-CO2</i>	0.0032	0.0028	0.0028	0.043	0.042	0.29	0.028	0.01
<i>MT-ATP8</i>	0.83	0.88	0.46	0.73	0.69	0.62	0.81	0.75
<i>MT-ATP6</i>	0.014	0.013	0.042	0.11	0.11	0.28	0.013	0.02
<i>MT-CO3</i>	0.95	0.96	0.53	0.83	0.88	0.58	0.74	0.86
<i>MT-ND3</i>	0.12	0.13	0.091	0.43	0.39	0.43	0.052	0.2
<i>MT-ND4L</i>	0.65	0.59	0.72	0.43	0.44	0.48	0.95	0.65
<i>MT-ND4</i>	0.12	0.15	0.057	0.09	0.099	0.39	0.68	0.21
<i>MT-ND5</i>	0.0015	0.0017	0.0041	0.066	0.00011	0.057	0.028	0.0024
<i>MT-ND6</i>	0.98	0.99	0.71	0.56	0.59	0.5	0.75	0.94
<i>MT-CYB</i>	0.38	0.4	0.25	0.19	0.16	0.35	0.41	0.41

We perform cohort-specific association analyses between heteroplasmic mutations and sex. Meta-analysis was performed with Fisher's method to combine the p-values. Burden, the original burden test; Burden-A, adaptive burden test; Burden-S, the z-score weighting burden test; Burden-V1, variable threshold burden test with minimum p; Burden-V2, variable threshold burden test with ACAT; SKAT, the sequence kernel association test; SKAT-O, the method combining the burden and SKAT; ACAT, the aggregated Cauchy association test combining the burden and SKAT. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.19 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 1 from the fixed-effect meta-analysis for participants of African American ancestry

Gene/region	BETA	SE	95% LCL	95% UCL	P
<i>D-loop</i>	0.036	0.024	-0.011	0.084	0.13
<i>MT-RNR1</i>	0.13	0.085	-0.034	0.3	0.12
<i>MT-RNR2</i>	0.066	0.067	-0.067	0.2	0.33
<i>MT-ND1</i>	0.094	0.081	-0.064	0.25	0.24
<i>MT-ND2</i>	0.017	0.11	-0.19	0.23	0.87
<i>MT-CO1</i>	0.12	0.06	-0.0015	0.23	0.053
<i>MT-CO2</i>	0.18	0.11	-0.041	0.4	0.11
<i>MT-ATP8</i>	0.086	0.2	-0.31	0.48	0.67
<i>MT-ATP6</i>	0.17	0.11	-0.045	0.38	0.12
<i>MT-CO3</i>	0.011	0.12	-0.23	0.25	0.93
<i>MT-ND3</i>	0.43	0.17	0.096	0.76	0.011
<i>MT-ND4L</i>	0.081	0.21	-0.34	0.5	0.7
<i>MT-ND4</i>	0.092	0.081	-0.068	0.25	0.26
<i>MT-ND5</i>	0.12	0.045	0.032	0.21	0.0076
<i>MT-ND6</i>	-0.022	0.13	-0.27	0.23	0.87
<i>MT-CYB</i>	0.1	0.073	-0.039	0.25	0.15

We perform cohort-specific association analyses between heteroplasmic mutations and sex. Meta-analysis was performed with the fixed-effects inverse variance method. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.20 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 2 from the fixed-effect meta-analysis for participants of African American ancestry

Gene/region	BETA	SE	95% LCL	95% UCL	P
<i>D-loop</i>	0.003	0.0012	0.00066	0.0053	0.012
<i>MT-RNR1</i>	0.0063	0.0032	-1.40E-05	0.013	0.051
<i>MT-RNR2</i>	0.003	0.0022	-0.0012	0.0073	0.16
<i>MT-ND1</i>	0.0024	0.0025	-0.0025	0.0073	0.33
<i>MT-ND2</i>	9.50E-05	0.0029	-0.0056	0.0058	0.97
<i>MT-CO1</i>	0.002	0.0019	-0.0017	0.0057	0.29
<i>MT-CO2</i>	0.0081	0.0035	0.0013	0.015	0.019
<i>MT-ATP8</i>	0.0021	0.0056	-0.0088	0.013	0.71
<i>MT-ATP6</i>	0.0086	0.0031	0.0026	0.015	0.0048
<i>MT-CO3</i>	-0.0015	0.0031	-0.0077	0.0047	0.63
<i>MT-ND3</i>	0.01	0.0052	-6.40E-05	0.02	0.051
<i>MT-ND4L</i>	-0.0023	0.0071	-0.016	0.012	0.75
<i>MT-ND4</i>	0.004	0.0027	-0.0013	0.0094	0.14
<i>MT-ND5</i>	0.0054	0.0016	0.0022	0.0086	0.00087
<i>MT-ND6</i>	-0.001	0.004	-0.0088	0.0068	0.8
<i>MT-CYB</i>	0.0035	0.0022	-0.00088	0.0078	0.12

We perform cohort-specific association analyses between heteroplasmic mutations and sex. Meta-analysis was performed with the fixed-effects inverse variance method. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.21 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 1 from Fisher's method meta-analysis for participants of European American ancestry

mtDNA region	P-values							
	Burden	Burden-A	Burden-S	Burden-V1	Burden-V2	SKAT	SKAT-O	ACAT
<i>D-loop</i>	0.95	0.86	0.94	0.27	0.31	0.45	0.81	0.79
<i>MT-RNR1</i>	0.47	0.47	0.53	0.15	0.47	0.71	0.74	0.66
<i>MT-RNR2</i>	0.66	0.66	0.65	0.58	0.97	0.46	0.57	0.62
<i>MT-ND1</i>	0.8	0.81	0.86	0.56	0.53	0.33	0.61	0.65
<i>MT-ND2</i>	0.51	0.65	0.76	0.12	0.18	0.6	0.61	0.58
<i>MT-CO1</i>	0.12	0.12	0.13	0.016	0.23	0.28	0.15	0.12
<i>MT-CO2</i>	0.24	0.23	0.19	0.042	0.034	0.29	0.27	0.25
<i>MT-ATP8</i>	0.66	0.67	0.66	0.22	0.21	0.38	0.74	0.54
<i>MT-ATP6</i>	0.24	0.24	0.21	0.11	0.12	0.42	0.27	0.28
<i>MT-CO3</i>	0.8	0.79	0.83	0.45	0.46	0.85	0.92	0.88
<i>MT-ND3</i>	0.33	0.23	0.18	0.86	0.87	0.5	0.38	0.37
<i>MT-ND4L</i>	0.53	0.56	0.67	0.83	0.84	0.59	0.56	0.54
<i>MT-ND4</i>	0.65	0.82	0.77	0.82	0.83	0.88	0.87	0.87
<i>MT-ND5</i>	0.38	0.11	0.087	0.29	0.83	0.82	0.68	0.56
<i>MT-ND6</i>	0.79	0.78	0.89	0.91	0.92	0.76	0.82	0.85
<i>MT-CYB</i>	0.69	0.67	0.71	0.19	0.095	0.19	0.48	0.43

We perform cohort-specific association analyses between heteroplasmic mutations and sex. Meta-analysis was performed with Fisher's method to combine the p-values. Burden, the original burden test; Burden-A, adaptive burden test; Burden-S, the z-score weighting burden test; Burden-V1, variable threshold burden test with minimum p; Burden-V2, variable threshold burden test with ACAT; SKAT, the sequence kernel association test; SKAT-O, the method combining the burden and SKAT; ACAT, the aggregated Cauchy association test combining the burden and SKAT. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.22 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 2 from Fisher's method meta-analysis for participants of European American ancestry

mtDNA region	P-values							
	Burden	Burden-A	Burden-S	Burden-V1	Burden-V2	SKAT	SKAT-O	ACAT
<i>D-loop</i>	0.81	0.59	0.77	0.075	0.85	0.49	0.48	0.74
<i>MT-RNR1</i>	0.34	0.34	0.31	0.024	0.46	0.46	0.78	0.38
<i>MT-RNR2</i>	0.4	0.41	0.32	0.61	1	0.62	0.95	0.52
<i>MT-ND1</i>	0.93	0.93	0.97	0.63	0.64	0.54	0.88	0.85
<i>MT-ND2</i>	0.65	0.65	0.72	0.24	0.21	0.58	0.89	0.71
<i>MT-CO1</i>	0.14	0.14	0.22	0.035	0.46	0.48	0.25	0.21
<i>MT-CO2</i>	0.14	0.14	0.1	0.17	0.17	0.46	0.14	0.21
<i>MT-ATP8</i>	0.75	0.74	0.75	0.22	0.22	0.54	0.62	0.68
<i>MT-ATP6</i>	0.19	0.19	0.15	0.19	0.17	0.54	0.31	0.31
<i>MT-CO3</i>	0.79	0.79	0.84	0.48	0.46	0.68	0.8	0.76
<i>MT-ND3</i>	0.39	0.39	0.33	0.94	0.94	0.8	0.87	0.6
<i>MT-ND4L</i>	0.16	0.16	0.25	0.78	0.79	0.57	0.7	0.2
<i>MT-ND4</i>	0.75	0.71	0.68	0.53	0.41	0.82	0.82	0.85
<i>MT-ND5</i>	0.18	0.072	0.096	0.34	0.12	0.62	0.38	0.28
<i>MT-ND6</i>	0.38	0.26	0.45	0.78	0.79	0.71	0.53	0.55
<i>MT-CYB</i>	0.54	0.53	0.51	0.24	0.36	0.5	0.79	0.52

We perform cohort-specific association analyses between heteroplasmic mutations and sex. Meta-analysis was performed with Fisher's method to combine the p-values. Burden, the original burden test; Burden-A, adaptive burden test; Burden-S, the z-score weighting burden test; Burden-V1, variable threshold burden test with minimum p; Burden-V2, variable threshold burden test with ACAT; SKAT, the sequence kernel association test; SKAT-O, the method combining the burden and SKAT; ACAT, the aggregated Cauchy association test combining the burden and SKAT. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.23 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 1 from the fixed-effect meta-analysis for participants of European American ancestry

Gene/region	BETA	SE	95% LCL	95% UCL	P
<i>D-loop</i>	-0.00031	0.019	-0.038	0.037	0.99
<i>MT-RNR1</i>	0.039	0.073	-0.1	0.18	0.6
<i>MT-RNR2</i>	0.047	0.058	-0.067	0.16	0.42
<i>MT-ND1</i>	-0.075	0.089	-0.25	0.099	0.4
<i>MT-ND2</i>	-0.12	0.076	-0.27	0.025	0.1
<i>MT-CO1</i>	0.028	0.062	-0.093	0.15	0.65
<i>MT-CO2</i>	0.16	0.1	-0.041	0.36	0.12
<i>MT-ATP8</i>	-0.14	0.18	-0.5	0.22	0.46
<i>MT-ATP6</i>	-0.0058	0.091	-0.19	0.17	0.95
<i>MT-CO3</i>	-0.047	0.091	-0.23	0.13	0.6
<i>MT-ND3</i>	0.072	0.14	-0.2	0.35	0.6
<i>MT-ND4L</i>	0.2	0.17	-0.13	0.54	0.24
<i>MT-ND4</i>	-0.036	0.061	-0.16	0.084	0.56
<i>MT-ND5</i>	0.047	0.044	-0.039	0.13	0.28
<i>MT-ND6</i>	-0.096	0.11	-0.3	0.11	0.37
<i>MT-CYB</i>	0.054	0.055	-0.053	0.16	0.33

We perform cohort-specific association analyses between heteroplasmic mutations and sex. Meta-analysis was performed with the fixed-effects inverse variance method. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.24 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 2 from the fixed-effect meta-analysis for participants of European American ancestry

Gene/region	BETA	SE	95% LCL	95% UCL	P
<i>D-loop</i>	0.00034	0.00081	-0.0012	0.0019	0.67
<i>MT-RNR1</i>	0.002	0.0016	-0.0012	0.0052	0.23
<i>MT-RNR2</i>	0.0016	0.0013	-0.0011	0.0042	0.24
<i>MT-ND1</i>	-0.00076	0.0019	-0.0045	0.003	0.69
<i>MT-ND2</i>	-0.0023	0.0018	-0.0057	0.0012	0.19
<i>MT-CO1</i>	0.0019	0.0014	-0.00075	0.0046	0.16
<i>MT-CO2</i>	0.0043	0.0023	-0.00012	0.0087	0.057
<i>MT-ATP8</i>	-0.003	0.0037	-0.01	0.0044	0.43
<i>MT-ATP6</i>	0.00029	0.002	-0.0036	0.0042	0.89
<i>MT-CO3</i>	5.00E-06	0.002	-0.0039	0.0039	1
<i>MT-ND3</i>	0.0029	0.0034	-0.0038	0.0096	0.39
<i>MT-ND4L</i>	0.0051	0.0036	-0.0018	0.012	0.15
<i>MT-ND4</i>	0.00068	0.0016	-0.0024	0.0037	0.66
<i>MT-ND5</i>	0.0018	0.0012	-0.00049	0.004	0.12
<i>MT-ND6</i>	0.00073	0.0024	-0.004	0.0055	0.76
<i>MT-CYB</i>	0.0023	0.0014	-0.00053	0.0051	0.11

We perform cohort-specific association analyses between heteroplasmic mutations and sex. Meta-analysis was performed with the fixed-effects inverse variance method. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.25 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 1 from Fisher's method meta-analysis for all participants

mtDNA region	P-values							
	Burden	Burden-A	Burden-S	Burden-V1	Burden-V2	SKAT	SKAT-O	ACAT
<i>D-loop</i>	0.62	0.5	0.47	0.32	0.36	0.3	0.49	0.53
<i>MT-RNR1</i>	0.5	0.48	0.15	0.28	0.64	0.74	0.73	0.69
<i>MT-RNR2</i>	0.74	0.74	0.49	0.53	0.75	0.31	0.54	0.57
<i>MT-ND1</i>	0.91	0.91	0.47	0.33	0.36	0.57	0.83	0.86
<i>MT-ND2</i>	0.45	0.59	0.45	0.085	0.14	0.78	0.65	0.59
<i>MT-CO1</i>	0.14	0.13	0.15	0.0062	0.22	0.075	0.062	0.059
<i>MT-CO2</i>	0.018	0.016	0.0083	0.064	0.054	0.31	0.04	0.042
<i>MT-ATP8</i>	0.82	0.86	0.41	0.51	0.47	0.65	0.92	0.78
<i>MT-ATP6</i>	0.066	0.064	0.11	0.17	0.19	0.5	0.099	0.097
<i>MT-CO3</i>	0.95	0.95	0.45	0.75	0.78	0.35	0.68	0.62
<i>MT-ND3</i>	0.13	0.095	0.047	0.56	0.55	0.38	0.19	0.24
<i>MT-ND4L</i>	0.56	0.54	0.53	0.71	0.7	0.39	0.46	0.45
<i>MT-ND4</i>	0.24	0.33	0.11	0.31	0.37	0.51	0.45	0.4
<i>MT-ND5</i>	0.0062	0.0012	0.0021	0.11	0.25	0.074	0.047	0.015
<i>MT-ND6</i>	0.96	0.96	0.95	0.84	0.83	0.67	0.89	0.94
<i>MT-CYB</i>	0.67	0.67	0.43	0.16	0.091	0.24	0.62	0.5

We perform cohort-specific association analyses between heteroplasmic mutations and age. Meta-analysis was performed with Fisher's method to combine the p-values in all participants. Burden, the original burden test; Burden-A, adaptive burden test; Burden-S, the z-score weighting burden test; Burden-V1, variable threshold burden test with minimum p; Burden-V2, variable threshold burden test with ACAT; SKAT, the sequence kernel association test; SKAT-O, the method combining the burden and SKAT; ACAT, the aggregated Cauchy association test combining the burden and SKAT. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.26 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 2 from Fisher's method meta-analysis for all participants

mtDNA region	P-values							
	Burden	Burden-A	Burden-S	Burden-V1	Burden-V2	SKAT	SKAT-O	ACAT
<i>D-loop</i>	0.22	0.14	0.15	0.17	0.8	0.38	0.19	0.26
<i>MT-RNR1</i>	0.31	0.3	0.16	0.059	0.53	0.57	0.25	0.46
<i>MT-RNR2</i>	0.34	0.35	0.24	0.34	0.54	0.36	0.38	0.38
<i>MT-ND1</i>	0.95	0.96	0.7	0.32	0.35	0.69	0.85	0.91
<i>MT-ND2</i>	0.49	0.55	0.44	0.11	0.12	0.74	0.86	0.67
<i>MT-CO1</i>	0.36	0.37	0.39	0.033	0.37	0.55	0.55	0.44
<i>MT-CO2</i>	0.0039	0.0035	0.0026	0.043	0.042	0.4	0.026	0.015
<i>MT-ATP8</i>	0.92	0.93	0.71	0.45	0.44	0.7	0.85	0.85
<i>MT-ATP6</i>	0.018	0.017	0.038	0.1	0.093	0.44	0.026	0.038
<i>MT-CO3</i>	0.97	0.97	0.81	0.77	0.77	0.76	0.9	0.93
<i>MT-ND3</i>	0.19	0.2	0.14	0.77	0.73	0.71	0.19	0.37
<i>MT-ND4L</i>	0.34	0.32	0.49	0.7	0.71	0.63	0.94	0.4
<i>MT-ND4</i>	0.31	0.35	0.16	0.19	0.17	0.68	0.88	0.49
<i>MT-ND5</i>	0.0025	0.0012	0.0035	0.11	0.00016	0.15	0.059	0.0056
<i>MT-ND6</i>	0.74	0.61	0.68	0.8	0.82	0.72	0.76	0.86
<i>MT-CYB</i>	0.53	0.54	0.39	0.19	0.22	0.48	0.69	0.54

We perform cohort-specific association analyses between heteroplasmic mutations and age. Meta-analysis was performed with Fisher's method to combine the p-values. Burden, the original burden test; Burden-A, adaptive burden test; Burden-S, the z-score weighting burden test; Burden-V1, variable threshold burden test with minimum p; Burden-V2, variable threshold burden test with ACAT; SKAT, the sequence kernel association test; SKAT-O, the method combining the burden and SKAT; ACAT, the aggregated Cauchy association test combining the burden and SKAT. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

Supplementary Table 2.27 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 1 from the fixed-effect meta-analysis of all participants

Gene/region	BETA	SE	95% LCL	95% UCL	P
<i>D-loop</i>	0.014	0.015	-0.016	0.043	0.36
<i>MT-RNR1</i>	0.078	0.055	-0.031	0.19	0.16
<i>MT-RNR2</i>	0.055	0.044	-0.031	0.14	0.21
<i>MT-ND1</i>	0.017	0.06	-0.1	0.13	0.77
<i>MT-ND2</i>	-0.076	0.063	-0.2	0.047	0.23
<i>MT-CO1</i>	0.076	0.043	-0.009	0.16	0.08
<i>MT-CO2</i>	0.17	0.074	0.024	0.31	0.022
<i>MT-ATP8</i>	-0.039	0.13	-0.3	0.22	0.77
<i>MT-ATP6</i>	0.066	0.07	-0.072	0.2	0.35
<i>MT-CO3</i>	-0.026	0.073	-0.17	0.12	0.72
<i>MT-ND3</i>	0.22	0.11	0.0049	0.43	0.045
<i>MT-ND4L</i>	0.15	0.13	-0.11	0.41	0.25
<i>MT-ND4</i>	0.01	0.049	-0.085	0.11	0.83
<i>MT-ND5</i>	0.083	0.031	0.021	0.14	0.0086
<i>MT-ND6</i>	-0.065	0.084	-0.23	0.099	0.44
<i>MT-CYB</i>	0.071	0.044	-0.015	0.16	0.11

We perform cohort-specific association analyses between heteroplasmic mutations and age. Meta-analysis was performed with the fixed-effects inverse variance method. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

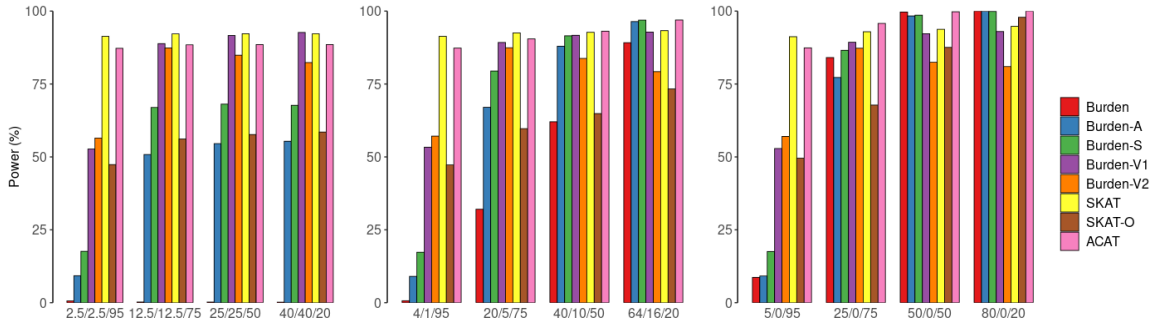
Supplementary Table 2.28 Association analyses between heteroplasmic variants of 16 mitochondrial genes/regions and sex for coding definition 2 from the fixed-effect meta-analysis of all participants

Gene/region	BETA	SE	95% LCL	95% UCL	P
<i>D-loop</i>	0.0012	0.00067	-0.00014	0.0025	0.081
<i>MT-RNR1</i>	0.0029	0.0014	5.50E-05	0.0057	0.046
<i>MT-RNR2</i>	0.002	0.0011	-0.00023	0.0042	0.08
<i>MT-ND1</i>	4.00E-04	0.0015	-0.0026	0.0034	0.79
<i>MT-ND2</i>	-0.0016	0.0015	-0.0046	0.0014	0.29
<i>MT-CO1</i>	0.0019	0.0011	-0.00027	0.0041	0.086
<i>MT-CO2</i>	0.0054	0.0019	0.0017	0.0092	0.0046
<i>MT-ATP8</i>	-0.0015	0.0031	-0.0075	0.0046	0.64
<i>MT-ATP6</i>	0.0027	0.0017	-0.00056	0.006	0.1
<i>MT-CO3</i>	-0.00044	0.0017	-0.0037	0.0029	0.79
<i>MT-ND3</i>	0.005	0.0028	-0.00055	0.011	0.077
<i>MT-ND4L</i>	0.0036	0.0032	-0.0027	0.0099	0.26
<i>MT-ND4</i>	0.0015	0.0014	-0.0012	0.0042	0.26
<i>MT-ND5</i>	0.0031	0.00096	0.0012	0.005	0.0013
<i>MT-ND6</i>	0.00027	0.0021	-0.0038	0.0043	0.89
<i>MT-CYB</i>	0.0026	0.0012	0.00033	0.005	0.025

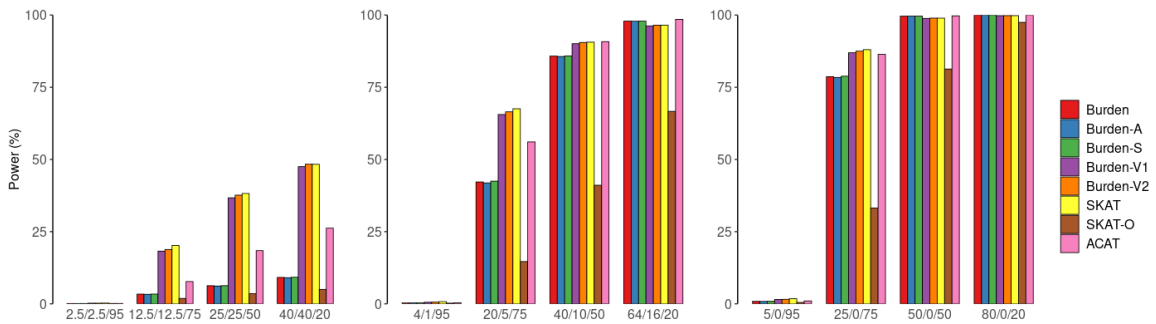
We perform cohort-specific association analyses between heteroplasmic mutations and age. Meta-analysis was performed with the fixed-effects inverse variance method. *MT-RNR1/RNR2*, the two ribosomal RNA genes in mitochondrial DNA; *MT-ND1/ND2/ND3/ND4/ND4L/ND5/ND6*, the mitochondrial NADH dehydrogenase, subunit 1, 2, 3, 4, 4L, 5 and 6 genes; *MT-CO1/CO2/CO3*, the mitochondrial cytochrome c oxidase I, II, and III genes; *MT-CYB*, the mitochondrial cytochrome b gene; *MT-APT6/ATP8*, the mitochondrial ATP synthase 6 and 8 genes.

A.2 Supplementary Figures

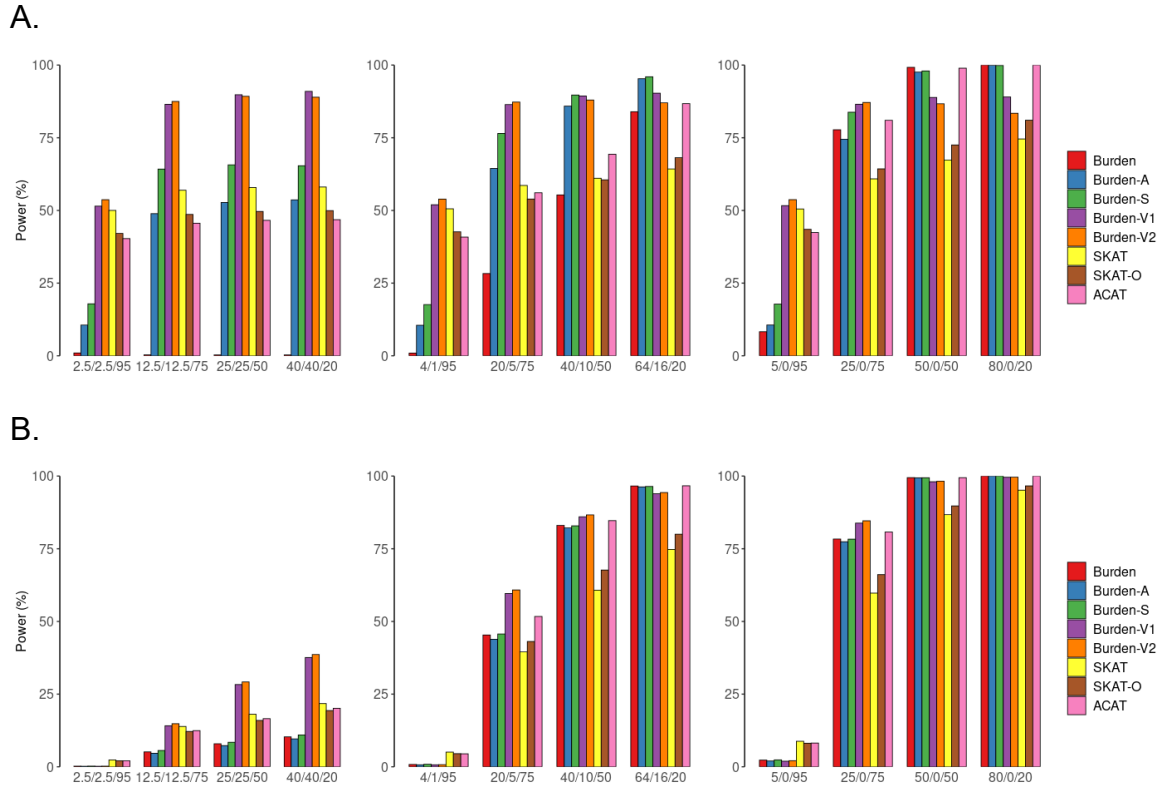
A.



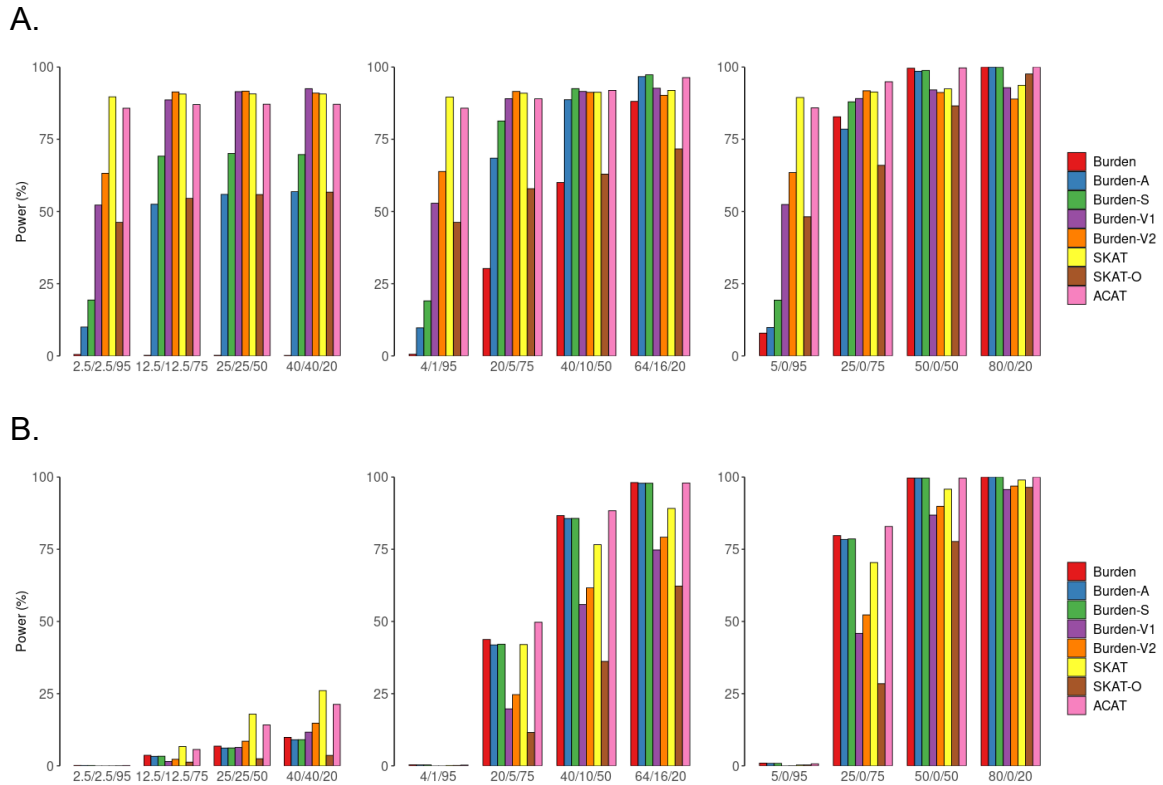
B.



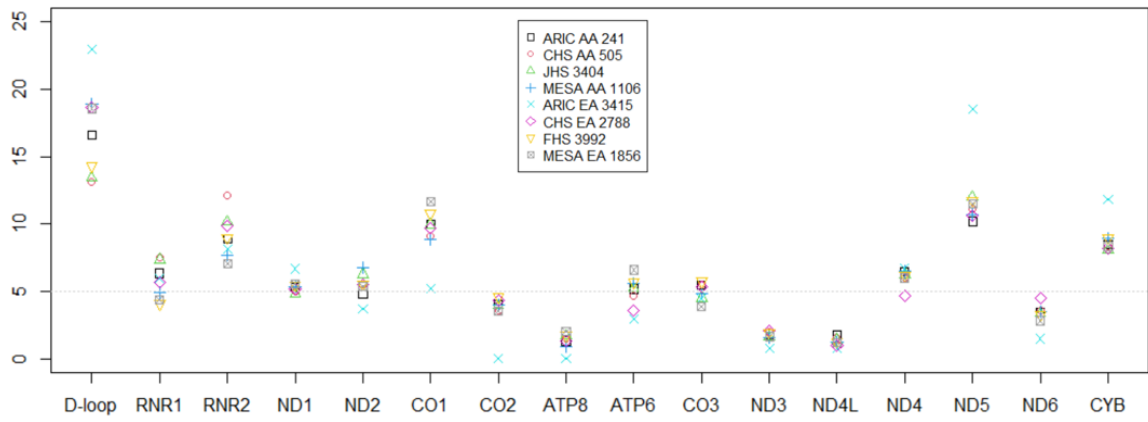
Supplementary Figure 2.1 Simulation study: power comparisons of six aggregate unit tests and two omnibus tests with a continuous trait and a binary trait for coding definition 2 (adjusted for empirical type I error rate). Power estimation was performed for a continuous trait (A) and a binary trait (B) at $\alpha=0.001$ with simulation data. Heteroplasmic mutations are defined by definition 2. We considered that 5%, 25%, 50%, or, 80% of the nonsynonymous heteroplasmic variants in the CYB gene are causal and consider that 50%, 80%, and 100% of the causal heteroplasmic mutations have effects with the same directionality. The variance that is explained by causal mutations is set to be 1% for the continuous trait and 2% for the binary trait. Burden, original burden test; Burden-A, adaptive burden test; Burden-S, z-score weighting burden test; Burden-V1, variable threshold burden test with minimum p-value; Burden-V2, variable threshold burden test with ACAT p-value combination method; SKAT, sequence kernel association test; SKAT-O, sequence kernel association test-optimal test; ACAT, aggregated Cauchy association test combining burden and SKAT. We simulated 50,000 replicates for evaluating power.



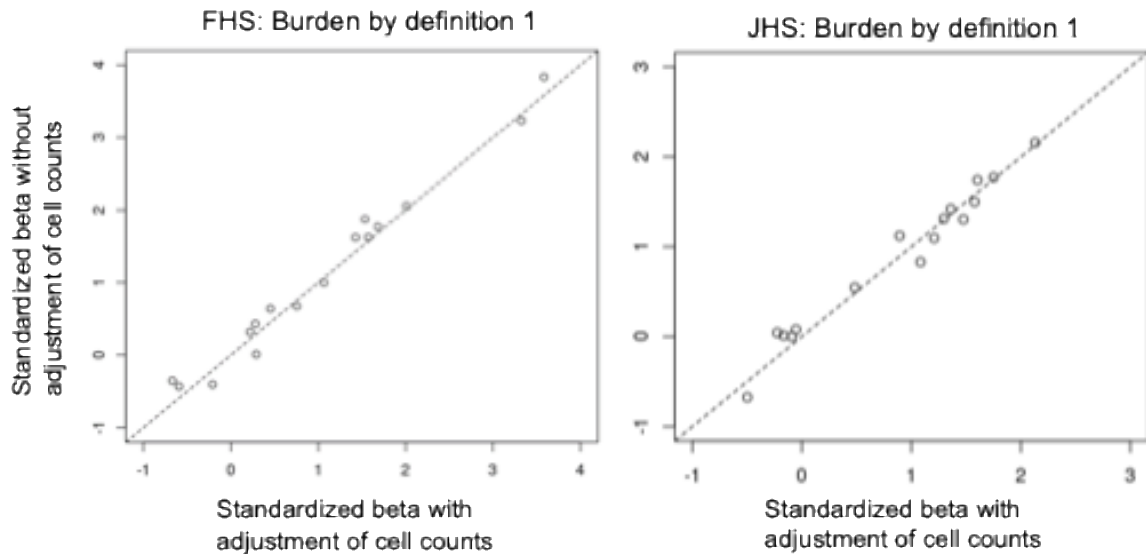
Supplementary Figure 2.2 Simulation-based power comparisons of six aggregate unit tests and two omnibus tests with a continuous and a binary trait for coding definition 1 (unadjusted for empirical type I error rate). Power estimation for a continuous trait (A) and a binary trait (B) at $\alpha=0.001$. Heteroplasmic variants are defined by an indicator function (definition 1). In simulations, we consider 5%, 25%, 50%, or 80% of the nonsynonymous heteroplasmic variants in the CYB gene to be causal and consider that 50%, 80%, and 100% of the causal heteroplasmic variants have effects with the same directionality. The variance that is explained by causal mutations is set to be 1% for the continuous trait and 2% for the binary trait. Burden, original burden test; Burden-A, adaptive burden test; Burden-S, z-score weighting burden test; Burden-V1, variable threshold burden test with minimum p-value; Burden-V2, variable threshold burden test with ACAT p-value combination method; SKAT, sequence kernel association test; SKAT-O, sequence kernel association test-optimal test; ACAT, aggregated Cauchy association test combining burden and SKAT. We simulated 50,000 replicates for evaluating the power.



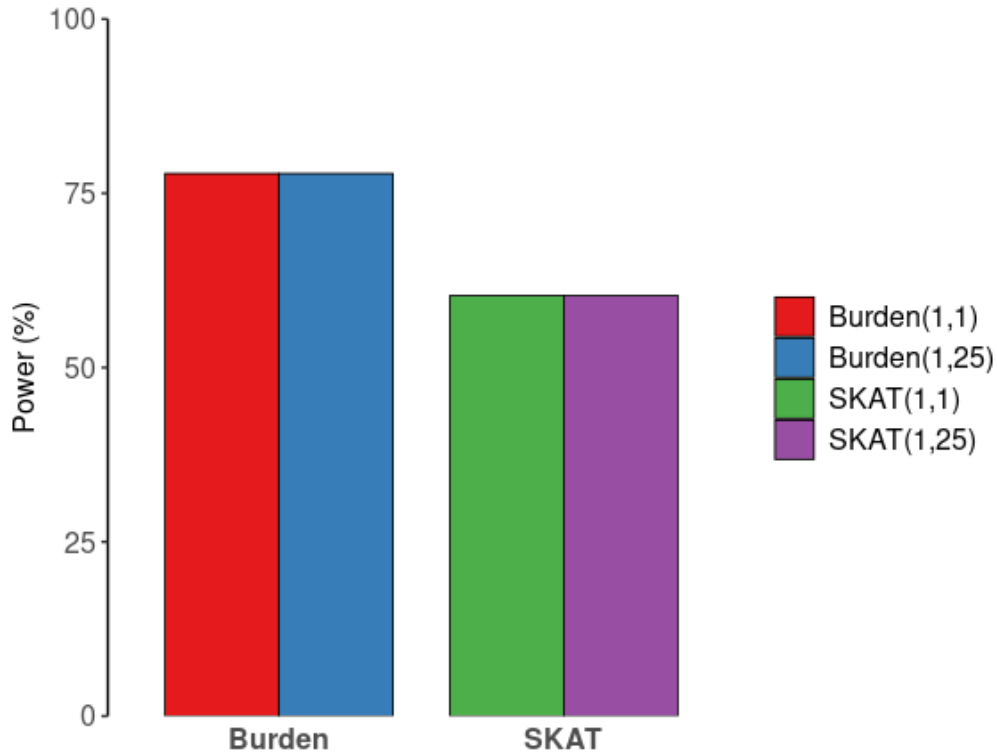
Supplementary Figure 2.3 Simulation study: power comparisons of six aggregate unit tests and two omnibus tests with a continuous trait and a binary trait for coding definition 2 (unadjusted for empirical type I error rate). Power estimation was performed for a continuous trait (A) and a binary trait (B) at $\alpha=0.001$ with simulation data. Heteroplasmic mutations are defined by definition 2. We considered that 5%, 25%, 50%, or, 80% of the nonsynonymous heteroplasmic variants in the CYB gene are causal and consider that 50%, 80%, and 100% of the causal heteroplasmic mutations have effects with the same directionality. The variance that is explained by causal mutations is set to be 1% for the continuous trait and 2% for the binary trait. Burden, original burden test; Burden-A, adaptive burden test; Burden-S, z-score weighting burden test; Burden-V1, variable threshold burden test with minimum p-value; Burden-V2, variable threshold burden test with ACAT p-value combination method; SKAT, sequence kernel association test; SKAT-O, sequence kernel association test-optimal test; ACAT, aggregated Cauchy association test combining burden and SKAT. We simulated 50,000 replicates for evaluating power.



Supplementary Figure 2.4 The proportion of heteroplasmic variants in each of the 15 genes and D-loop region



Supplementary Figure 2.5 Comparison of standardized beta coefficients of simple burden test of coding definition 1 with/o cell count variables in Framingham Heart Study (FHS) and Jackson Heart Study (JHS). We investigated whether adjusting for cell count and compositions affect the association of heteroplasmic burden in the 16 genes/regions and age in the same participants in the FHS (n = 2551) and JHS (n=2737).



Supplementary Figure 2.6 Simulation study: power comparisons at beta (MAF_H , 1, 1) and beta (MAF_H , 1, 25). Power estimation was performed for a continuous trait at $\alpha=0.001$ with simulation data. Heteroplasmic mutations are defined by definition 1. We considered that 25% of the nonsynonymous heteroplasmic variants in the CYB gene are causal and considered that 100% of the causal heteroplasmic mutations have effects with the same directionality. The variance that is explained by causal mutations is set to be 1%. Burden (1,1), the original burden test with weights of beta (MAF_H , 1, 1); Burden(1,25), the original burden test with weights of beta (MAF_H , 1, 25); SKAT(1,1), sequence kernel association test with weights of beta (MAF_H , 1, 1); SKAT (1,25), sequence kernel association test with weights of beta (MAF_H , 1, 25). We simulate 50,000 replicates for evaluating type I error.

APPENDIX B: SUPPLEMENTARY MATERIALS FOR CHAPTER 3

B.1 Supplementary Tables

Supplementary Table 3.1 MSE of estimated clusters' means for a continuous trait with the presence of LD in simulation studies

Scenario	N=5000	N=15000	N=25000
Scenario 1	0.0102	0.00429	0.00225
Scenario 2	0.0118	0.00226	7.00E-04
Scenario 3	0.014	0.00446	0.0022
Scenario 4	0.0148	0.00557	0.00268
Scenario 5	0.00405	0.00282	0.00177
Scenario 6	0.0061	0.00042	0.000116

Supplementary Table 3.2 Accuracy of number of clusters specification for a continuous trait with the absence of LD in simulation studies

Scenario	N=5000	N=15000	N=25000
Scenario 1	0.132	0.911	0.98
Scenario 2	0.99	0.995	0.991
Scenario 3	0.093	0.994	0.995
Scenario 4	0.001	0.691	0.993
Scenario 5	0.737	0.895	0.935
Scenario 6	0.788	0.91	0.954

Supplementary Table 3.3 Accuracy of number of clusters specification for a continuous trait with the presence of LD in simulation studies

Scenario	N=5000	N=15000	N=25000
Scenario 1	0.034	0.715	0.946
Scenario 2	0.995	0.992	0.993
Scenario 3	0.031	0.965	0.995
Scenario 4	0	0.596	0.977
Scenario 5	0.585	0.828	0.874
Scenario 6	0.789	0.906	0.934

Supplementary Table 3.4 Accuracy of number of clusters specification for a binary trait with the absence of LD in simulation studies

Scenario	N=5000	N=15000	N=25000
Scenario 1	0.004	0.258	0.711
Scenario 2	0.56	0.997	0.998
Scenario 3	0.003	0.152	0.77
Scenario 4	0	0	0.057
Scenario 5	0.476	0.761	0.842
Scenario 6	0.737	0.792	0.86

Supplementary Table 3.5 Accuracy of number of clusters specification for a binary trait with the presence of LD in simulation studies

Scenario	N=5000	N=15000	N=25000
Scenario 1	0.002	0.069	0.292
Scenario 2	0.47	0.99	0.996
Scenario 3	0	0.078	0.567
Scenario 4	0	0.001	0.038
Scenario 5	0.424	0.647	0.766
Scenario 6	0.748	0.784	0.854

Supplementary Table 3.6 Summary of associations between genes and blood pressure traits in real data analysis

	Chr	Start pos	End pos	Number of rare variants	P-value		
					SBP	DBP	PP
CEP120	5	122682248	122758665	21	-	-	8.25E-31
COL21A1	6	55923968	56047400	26	-	-	2.3E-29
DBH	9	136501569	136523555	29	8.85E-22	1.77E-32	-
NOX4	11	89069094	89224387	9	-	-	1.85E-23
NPR1	1	153652129	153665650	13	2.72E-10	-	-
PLCB3	11	64021930	64034975	15	4.97E-13	4.27E-10	-

Supplementary Table 3.7 Summary of clustering results for rare variants within signal genes of DBP

Gene	Chr	Start (bp)	End (bp)	# variants	# clusters	mu	phi	# variants of clusters	P-value _ANOVA	P-value MWU
DBH	9	136501569	136523555	27	2	0/-0.0927	0.786/0.214	22/5	9.97e-07	0.0391
PLCB3	11	64021930	64034975	18	2	0/0.0473	0.69/0.31	14/4	0.000166	0.626

Supplementary Table 3.8 Summary of clustering results for rare variants within signal genes of PP

Gene	Chr	Start (bp)	End (bp)	# variants	# clusters	mu	phi	# variants of clusters	P-value _ANOVA	P-value MWU
CEP120	5	122682248	122758665	21	2	0/-0.0435	0.673/0.327	18/3	9.98e-08	0.196
COL21A1	6	55923968	56047400	24	3	0/0.335/0.171	0.761/0.0493/0.19	19/1/4	1.14e-09	0.312
NOX4	11	89069094	89224387	10	2	0/-0.0962	0.605/0.395	7/3	0.00125	0.937

Supplementary Table 3.9 Summary of clustering results for rare variants within signal genes of HTN

Gene	Chr	Start (bp)	End (bp)	# variants	# clusters	mu	phi	# variants of clusters	P-value _ANOVA	P-value MWU
DBH	9	136501569	136523555	29	2	0/-2.612	0.712/0.288	21/8	2.15e-07	0.495
NPR1	1	153652129	153665650	12	3	0/-2.053/4.334	0.667/0.253/0.079	8/3/1	0.000282	0.461
PLCB3	11	64021930	64034975	18	2	0/4.808	0.887/0.113	16/2	1.27e-05	0.433

Supplementary Table 3.10 Clustering results of rare variants within the combined signal region of SBP

rsID	Chr	Position	Major allele	Minor allele	MAF	Effect	SE	P-value	Cluster membership	Cluster mean
rs76856960	9	136501569	g	a	0.0046	0.00322	0.0138	0.8151	1	0
rs143544421	9	136501599	g	a	5e-04	0.123	0.156	0.4303	1	0
rs78445536	9	136501746	g	a	5e-04	-0.0666	0.0572	0.2443	1	0
rs146922432	9	136501768	g	a	3e-04	-0.0174	0.144	0.9042	1	0
rs76819676	9	136507375	g	a	2e-04	-0.0232	0.111	0.8347	1	0
rs200430427	9	136507456	c	t	4e-04	-0.0581	0.154	0.706	1	0
rs143535251	9	136507474	c	t	0.0013	0.0197	0.0248	0.4267	1	0
rs200628504	9	136507528	c	t	4e-04	0.0516	0.128	0.6857	1	0
rs5321	9	136507559	g	c	5e-04	0.0625	0.069	0.3652	1	0
rs199734841	9	136507586	g	c	4e-04	-0.176	0.2	0.3791	1	0
rs13306301	9	136508640	g	a	1e-04	-0.549	0.282	0.05178	1	0
rs5324	9	136508658	g	a	6e-04	-0.0219	0.0384	0.5682	1	0
rs145655199	9	136508682	g	a	0.001	-0.0173	0.18	0.9233	1	0
rs201681337	9	136508691	g	a	0.002	0.0305	0.0754	0.6864	1	0
rs75215331	9	136513028	c	t	0.0034	0.0146	0.0168	0.3834	1	0
rs41316996	9	136521654	g	a	0.0032	0.019	0.0145	0.1918	1	0
rs144040856	9	136521738	g	a	3e-04	-0.0756	0.135	0.5758	1	0
rs141021210	9	136521751	a	g	0	-0.131	0.997	0.8956	1	0
rs201973877	9	136522317	t	c	0.0003	-0.0251	0.0565	0.6566	1	0
rs75512464	9	136523487	a	t	0.002	-0.0264	0.0236	0.2621	1	0
rs76316834	9	136523555	g	a	4e-04	0.0484	0.13	0.709	1	0
rs56019647	1	153652129	c	t	1e-04	-0.113	0.165	0.4947	1	0
rs28730726	1	153653757	g	c	0.0016	-0.0175	0.0432	0.6842	1	0
rs199612927	1	153654234	g	a	2e-04	0.0688	0.0871	0.4298	1	0
rs140425746	1	153655966	g	a	2e-04	-0.0333	0.154	0.8287	1	0
rs201746049	1	153656216	a	g	0.0003	-0.033	0.195	0.8658	1	0
rs115938602	1	153656228	c	a	0	-0.139	0.29	0.6323	1	0
rs149202797	1	153658306	c	a	2e-04	-0.0203	0.136	0.8808	1	0
rs116775696	1	153659550	a	g	0	-0.333	0.458	0.4672	1	0

rs139174442	1	153660154	c	g	10.0e-05	0.297	0.35	0.3973	1	0
rs201787421	1	153660625	g	a	2e-04	-0.68	0.316	0.0312	1	0
rs200996360	11	640253	c	g	0.0003	-0.03	0.0527	0.5689	1	0
rs111961110	11	64021930	a	g	0.0003	-0.14	0.207	0.4974	1	0
rs144191345	11	64022785	c	t	2e-04	-0.0212	0.108	0.8442	1	0
rs199645363	11	64022873	c	t	3e-04	0.0912	0.101	0.3668	1	0
rs138400940	11	64022905	c	t	1e-04	-0.381	0.576	0.5084	1	0
rs188572550	11	64026357	g	a	0	0.801	0.704	0.2551	1	0
rs201842672	11	64027567	g	t	0.0038	-0.0627	0.0711	0.3777	1	0
rs200923408	11	64028916	g	a	3e-04	-0.538	0.235	0.02196	1	0
rs200263631	11	64031571	t	c	0.0003	0.0847	0.0703	0.2283	1	0
rs79573066	11	64032784	g	a	0.0036	0.0262	0.0696	0.7066	1	0
rs61757725	11	64033360	g	t	0.009	0.0786	0.0599	0.1896	1	0
rs148059922	11	64033391	g	c	4e-04	-0.0149	0.131	0.9093	1	0
rs201342752	11	64033986	c	t	0.002	-0.0891	0.0706	0.2068	1	0
rs141163685	11	64034734	g	t	0.0031	0.249	0.113	0.02777	1	0
rs113554478	11	64034975	g	a	0.0011	0.00854	0.102	0.9333	1	0
rs3025380	9	136501756	g	c	0.0045	-0.0884	0.0119	1.096e-13	2	-0.0826
rs74853476	9	136501834	t	c	0.0021	-0.0774	0.0181	1.973e-05	2	-0.0826
rs142383279	9	136507332	g	a	0.0018	-0.0652	0.0194	0.0007811	2	-0.0826
rs145059403	9	136507425	g	a	0.001	-0.102	0.0261	9.144e-05	2	-0.0826
rs148439785	9	136521726	g	a	8e-04	-0.0595	0.0313	0.05706	2	-0.0826
rs151228388	9	136522272	a	g	0.0013	-0.317	0.102	0.001909	2	-0.0826
rs61757359	1	153658297	g	a	0.0034	-0.0819	0.014	4.487e-09	2	-0.0826
rs61758562	1	153659131	g	a	5e-04	-0.133	0.0495	0.007185	2	-0.0826
rs116245325	1	153665650	c	t	8e-04	0.166	0.0287	7.686e-09	3	0.0573
rs117874826	11	64027666	a	c	0.0138	0.0467	0.00729	1.535e-10	3	0.0573
rs145502455	11	64031030	g	a	0.0054	0.0723	0.0118	8.563e-10	3	0.0573
rs142330950	11	64032945	g	t	0.0021	0.0456	0.0182	0.01236	3	0.0573

Supplementary Table 3.11 Clustering results of rare variants within the combined signal region of DBP

rsID	Chr	Position	Major allele	Minor allele	MAF	Effect	SE	P-value	Cluster membership	Cluster mean
rs76856960	9	136501569	g	a	0.0046	-0.00876	0.0138	0.5248	1	0
rs143544421	9	136501599	g	a	5e-04	0.0145	0.156	0.9262	1	0
rs78445536	9	136501746	g	a	5e-04	-0.0651	0.0572	0.255	1	0
rs146922432	9	136501768	g	a	3e-04	-0.125	0.144	0.3888	1	0
rs76819676	9	136507375	g	a	2e-04	0.0357	0.111	0.7477	1	0
rs200430427	9	136507456	c	t	4e-04	0.0385	0.154	0.8029	1	0
rs143535251	9	136507474	c	t	0.0013	0.00881	0.0248	0.7226	1	0
rs200628504	9	136507528	c	t	4e-04	-0.081	0.128	0.5253	1	0
rs5321	9	136507559	g	c	5e-04	0.0583	0.069	0.3984	1	0
rs199734841	9	136507586	g	c	4e-04	-0.225	0.2	0.2601	1	0
rs13306301	9	136508640	g	a	1e-04	-0.753	0.283	0.007722	1	0
rs5324	9	136508658	g	a	6e-04	-0.0431	0.0385	0.2625	1	0
rs145655199	9	136508682	g	a	0.001	0.129	0.18	0.473	1	0
rs201681337	9	136508691	g	a	0.002	0.0397	0.0755	0.5993	1	0
rs75215331	9	136513028	c	t	0.0034	0.00505	0.0168	0.7633	1	0
rs41316996	9	136521654	g	a	0.0032	0.00855	0.0146	0.557	1	0
rs148439785	9	136521726	g	a	8e-04	-0.0323	0.0313	0.3019	1	0
rs144040856	9	136521738	g	a	3e-04	-0.217	0.135	0.1089	1	0
rs141021210	9	136521751	a	g	0	-1.023	0.995	0.3037	1	0
rs201973877	9	136522317	t	c	0.0003	-0.0286	0.0566	0.6134	1	0
rs75512464	9	136523487	a	t	0.002	-0.025	0.0236	0.2893	1	0
rs76316834	9	136523555	g	a	4e-04	0.015	0.13	0.9078	1	0
rs200996360	11	640253	c	g	0.0003	0.0237	0.0528	0.653	1	0
rs111961110	11	64021930	a	g	0.0003	0.00639	0.207	0.9754	1	0
rs144191345	11	64022785	c	t	2e-04	0.0759	0.108	0.4822	1	0
rs199645363	11	64022873	c	t	3e-04	0.0863	0.101	0.3935	1	0

rs138400940	11	64022905	c	t	1e-04	-0.858	0.576	0.1362	1	0
rs188572550	11	64026357	g	a	0	0.624	0.704	0.3759	1	0
rs201842672	11	64027567	g	t	0.0038	-0.0781	0.0711	0.2721	1	0
rs200923408	11	64028916	g	a	3e-04	-0.194	0.235	0.4077	1	0
rs200263631	11	64031571	t	c	0.0003	0.121	0.0704	0.08613	1	0
rs79573066	11	64032784	g	a	0.0036	0.0198	0.0697	0.7768	1	0
rs61757725	11	64033360	g	t	0.009	0.0537	0.06	0.3709	1	0
rs148059922	11	64033391	g	c	4e-04	-0.154	0.131	0.2409	1	0
rs201342752	11	64033986	c	t	0.002	-0.0181	0.0706	0.7979	1	0
rs141163685	11	64034734	g	t	0.0031	0.214	0.113	0.05838	1	0
rs113554478	11	64034975	g	a	0.0011	0.0951	0.102	0.3518	1	0
rs3025380	9	136501756	g	c	0.0045	-0.103	0.0119	4.287e-18	2	-0.0924
rs74853476	9	136501834	t	c	0.0021	-0.0954	0.0182	1.529e-07	2	-0.0924
rs142383279	9	136507332	g	a	0.0018	-0.0701	0.0194	0.0003159	2	-0.0924
rs145059403	9	136507425	g	a	0.001	-0.0829	0.0262	0.001551	2	-0.0924
rs151228388	9	136522272	a	g	0.0013	-0.226	0.102	0.02711	2	-0.0924
rs117874826	11	64027666	a	c	0.0138	0.041	0.00731	2.069e-08	3	0.0471
rs145502455	11	64031030	g	a	0.0054	0.0668	0.0118	1.619e-08	3	0.0471
rs142330950	11	64032945	g	t	0.0021	0.0367	0.0182	0.04445	3	0.0471

Supplementary Table 3.12 Clustering results of rare variants within the combined signal region of PP

rsID	Chr	Position	Major allele	Minor allele	MAF	Effect	SE	P-value	Cluster membership	Cluster mean
rs145436175	5	122682248	t	c	0.0008	0.233	0.268	0.385	1	0
rs140306974	5	122685717	g	a	0.0011	0.0594	0.126	0.6367	1	0
rs200061679	5	122685731	t	a	0.0015	0.0692	0.0388	0.07491	1	0
rs142792779	5	122708381	c	t	3e-04	-0.0136	0.213	0.9492	1	0
rs139865050	5	122713092	c	g	0	-0.192	0.707	0.7861	1	0
rs74938108	5	122713159	c	t	0.0016	-0.0257	0.0905	0.7765	1	0
rs201600892	5	122713191	c	g	10.0e-05	0.0767	0.447	0.8637	1	0
rs61744334	5	122714044	t	c	0.0006	-0.0201	0.0684	0.7689	1	0
rs144490830	5	122714104	g	c	2e-04	0.37	0.179	0.03908	1	0
rs147277049	5	122720724	t	c	0.0006	-0.0181	0.036	0.6152	1	0
rs200450605	5	122725693	g	a	0.0017	0.016	0.0229	0.4841	1	0
rs201571160	5	122725754	c	g	0.0003	0.166	0.316	0.6004	1	0
rs114281792	5	122725761	t	c	0.0067	-0.00233	0.0098	0.8124	1	0
rs61747983	5	122725768	g	a	1e-04	0.0156	0.112	0.8895	1	0
rs147273517	5	122748194	t	c	10.0e-05	0.36	0.162	0.02599	1	0
rs202103949	5	122748198	t	c	0.0004	-0.029	0.121	0.8106	1	0
rs199793672	5	122758665	t	c	10.0e-05	-0.479	0.503	0.3404	1	0
rs200478915	6	55923968	g	c	8e-04	-0.0449	0.0323	0.1652	1	0
rs200564236	6	55925008	g	a	8e-04	-0.0212	0.162	0.8957	1	0
rs199722485	6	55925588	t	g	0.0005	-0.0128	0.106	0.9038	1	0
rs200674177	6	55925689	g	a	2e-04	-0.0981	0.189	0.6046	1	0
rs201839603	6	55925762	a	c	10.0e-05	0.0193	0.577	0.9733	1	0
rs9464337	6	55925801	g	t	0.0054	0.000204	0.0116	0.986	1	0
rs201892311	6	55926464	g	c	1e-04	0.426	0.707	0.547	1	0
rs199910287	6	55935556	g	a	6e-04	0.054	0.0445	0.2255	1	0
rs191626317	6	55939059	g	t	0.002	0.0424	0.102	0.6783	1	0

rs75605879	6	55966311	a	g	10.0e-05	-0.147	0.107	0.169	1	0
rs202115077	6	55988871	g	t	0.002	0.00866	0.0186	0.6409	1	0
rs201267383	6	56006611	t	c	0.0003	-0.077	0.0817	0.3461	1	0
rs35583895	6	56006732	a	g	0.0081	0.00138	0.0479	0.9771	1	0
rs200361985	6	56029264	g	a	0.002	0.0616	0.0546	0.2593	1	0
rs142653960	6	56035494	c	t	4e-04	-0.0656	0.0525	0.2115	1	0
rs199532612	6	56035853	a	t	0.0017	-0.0186	0.0217	0.3925	1	0
rs200708113	6	56035881	c	t	5e-04	-0.0112	0.0509	0.8259	1	0
rs202026963	6	56035909	g	a	1e-04	0.0193	0.242	0.9364	1	0
rs147394600	6	56047400	g	a	1e-04	-0.159	0.164	0.331	1	0
rs115031759	11	89073269	g	a	1e-04	0.192	0.161	0.2323	1	0
rs201165492	11	89106599	c	t	1e-04	-0.274	0.576	0.6345	1	0
rs149515506	11	89135492	a	g	10.0e-05	0.428	0.706	0.5445	1	0
rs147350656	11	89177310	c	t	1e-04	-0.184	0.236	0.4375	1	0
rs142433357	11	89182609	t	c	10.0e-05	-1.231	0.707	0.08181	1	0
rs55977241	11	89182652	c	t	0.002	0.000803	0.109	0.9941	1	0
rs145686545	11	89224387	c	t	2e-04	0.162	0.202	0.4229	1	0
rs2303720	5	122682334	c	t	0.0291	-0.0418	0.00482	4.532e-18	2	-0.0438
rs114280473	5	122714092	g	a	0.0063	-0.0632	0.011	8.076e-09	2	-0.0438
rs189429890	5	122729025	c	t	0.0065	-0.0373	0.00999	0.0001886	2	-0.0438
rs144215891	11	89069094	t	c	0.0014	-0.0716	0.0302	0.01798	2	-0.0438
rs139341533	11	89182666	c	a	0.0043	-0.0905	0.0126	8.085e-13	3	-0.0968
rs56061986	11	89182686	t	c	0.0029	-0.113	0.0161	2.385e-12	3	-0.0968
rs200999181	6	55935568	c	a	0.0012	0.336	0.0244	3.471e-43	4	0.334
rs201955087	5	122727015	c	t	0.0013	0.274	0.106	0.009927	5	0.173
rs115079907	6	55924005	c	t	0.0015	0.207	0.0248	5.573e-17	5	0.173
rs76146749	6	55925783	t	a	7e-04	0.198	0.0782	0.01144	5	0.173
rs200401514	6	55989091	c	t	4e-04	0.124	0.0552	0.02532	5	0.173
rs2764043	6	56035643	a	g	0.0016	0.152	0.0203	7.82e-14	5	0.173

Supplementary Table 3.13 Clustering results of rare variants within the combined signal region of HTN

rsID	Chr	Position	Major allele	Minor allele	MAF	Z value	P-value	Cluster membership	Cluster mean
rs76856960	9	136501569	g	a	0.0035	-0.511	0.6096	1	0
rs143544421	9	136501599	g	a	2e-04	-0.526	0.5989	1	0
rs78445536	9	136501746	g	a	3e-04	0.234	0.815	1	0
rs146922432	9	136501768	g	a	1e-04	1.257	0.2089	1	0
rs142383279	9	136507332	g	a	0.002	-1.826	0.06781	1	0
rs76819676	9	136507375	g	a	1e-04	-0.787	0.4313	1	0
rs143535251	9	136507474	c	t	0.0014	0.397	0.6916	1	0
rs200628504	9	136507528	c	t	2e-04	-0.0675	0.9462	1	0
rs5321	9	136507559	g	c	2e-04	0.226	0.8212	1	0
rs199734841	9	136507586	g	c	2e-04	0.897	0.3697	1	0
rs13306301	9	136508640	g	a	0	-1.378	0.1683	1	0
rs5324	9	136508658	g	a	8e-04	-1.404	0.1604	1	0
rs145655199	9	136508682	g	a	3e-04	-1.247	0.2125	1	0
rs201681337	9	136508691	g	a	2e-04	1.109	0.2676	1	0
rs75215331	9	136513028	c	t	0.0025	-0.0818	0.9348	1	0
rs41316996	9	136521654	g	a	0.003	2.028	0.04261	1	0
rs144040856	9	136521738	g	a	1e-04	0.0598	0.9523	1	0
rs141021210	9	136521751	a	g	0	-0.665	0.5059	1	0
rs201973877	9	136522317	t	c	0.0002	-0.923	0.356	1	0
rs75512464	9	136523487	a	t	0.0014	-0.289	0.7723	1	0
rs148806316	9	136523534	c	t	0.0025	0.295	0.7679	1	0
rs76316834	9	136523555	g	a	2e-04	-0.39	0.6962	1	0
rs56019647	1	153652129	c	t	0	-0.727	0.4675	1	0
rs199612927	1	153654234	g	a	1e-04	0.468	0.6397	1	0
rs140425746	1	153655966	g	a	1e-04	-1.607	0.1081	1	0
rs201746049	1	153656216	a	g	10.0e-05	-1.254	0.21	1	0
rs115938602	1	153656228	c	a	0	-1.174	0.2405	1	0
rs149202797	1	153658306	c	a	1e-04	0.23	0.8182	1	0

rs61758562	1	153659131	g	a	0.0011	-1.693	0.09054	1	0
rs116775696	1	153659550	a	g	0	0.641	0.5213	1	0
rs139174442	1	153660154	c	g	0	0.0493	0.9606	1	0
rs201787421	1	153660625	g	a	1e-04	-0.287	0.7739	1	0
rs200996360	11	640253	c	g	0.0003	-0.905	0.3653	1	0
rs111961110	11	64021930	a	g	10.0e-05	-0.917	0.3591	1	0
rs144191345	11	64022785	c	t	1e-04	0.259	0.7958	1	0
rs199645363	11	64022873	c	t	2e-04	-0.587	0.5575	1	0
rs138400940	11	64022905	c	t	1e-04	-1.28	0.2004	1	0
rs188572550	11	64026357	g	a	0	-0.324	0.7457	1	0
rs201842672	11	64027567	g	t	9e-04	-0.58	0.5616	1	0
rs200923408	11	64028916	g	a	1e-04	-0.496	0.6198	1	0
rs200263631	11	64031571	t	c	0.0002	1.033	0.3018	1	0
rs79573066	11	64032784	g	a	2e-04	0.665	0.5063	1	0
rs142330950	11	64032945	g	t	0.0021	2.175	0.02966	1	0
rs61757725	11	64033360	g	t	3e-04	0.427	0.6694	1	0
rs148059922	11	64033391	g	c	1e-04	1.308	0.1908	1	0
rs201342752	11	64033986	c	t	5e-04	1.223	0.2215	1	0
rs141163685	11	64034734	g	t	4e-04	0.867	0.3859	1	0
rs113554478	11	64034975	g	a	1e-04	0.708	0.4787	1	0
rs77273740	9	136501728	c	t	0.0042	-2.933	0.003353	2	-2.785
rs3025380	9	136501756	g	c	0.0045	-5.297	1.176e-07	2	-2.785
rs74853476	9	136501834	t	c	0.0022	-2.186	0.02885	2	-2.785
rs145059403	9	136507425	g	a	0.0012	-3.468	0.0005248	2	-2.785
rs200430427	9	136507456	c	t	1e-04	-2.306	0.02112	2	-2.785
rs148439785	9	136521726	g	a	0.0019	-2.196	0.02806	2	-2.785
rs151228388	9	136522272	a	g	0.0003	-2.399	0.01646	2	-2.785
rs61757359	1	153658297	g	a	0.0036	-3.719	2e-04	2	-2.785
rs116245325	1	153665650	c	t	8e-04	4.335	1.457e-05	3	4.646
rs117874826	11	64027666	a	c	0.0308	5.712	1.119e-08	3	4.646
rs145502455	11	64031030	g	a	0.0047	4.015	5.952e-05	3	4.646

Supplementary Table 3.14 Clustering results of rare variants within the DBH gene with SBP

rsID	Chr	Position	Major allele	Minor allele	MAF	Effect	SE	P-value	Cluster membership	Cluster mean
rs76856960	9	136501569	g	a	0.0046	0.00322	0.0138	0.8151	1	0
rs143544421	9	136501599	g	a	5e-04	0.123	0.156	0.4303	1	0
rs78445536	9	136501746	g	a	5e-04	-0.0666	0.0572	0.2443	1	0
rs146922432	9	136501768	g	a	3e-04	-0.0174	0.144	0.9042	1	0
rs76819676	9	136507375	g	a	2e-04	-0.0232	0.111	0.8347	1	0
rs200430427	9	136507456	c	t	4e-04	-0.0581	0.154	0.706	1	0
rs143535251	9	136507474	c	t	0.0013	0.0197	0.0248	0.4267	1	0
rs200628504	9	136507528	c	t	4e-04	0.0516	0.128	0.6857	1	0
rs5321	9	136507559	g	c	5e-04	0.0625	0.069	0.3652	1	0
rs199734841	9	136507586	g	c	4e-04	-0.176	0.2	0.3791	1	0
rs13306301	9	136508640	g	a	1e-04	-0.549	0.282	0.05178	1	0
rs5324	9	136508658	g	a	6e-04	-0.0219	0.0384	0.5682	1	0
rs145655199	9	136508682	g	a	0.001	-0.0173	0.18	0.9233	1	0
rs201681337	9	136508691	g	a	0.002	0.0305	0.0754	0.6864	1	0
rs75215331	9	136513028	c	t	0.0034	0.0146	0.0168	0.3834	1	0
rs41316996	9	136521654	g	a	0.0032	0.019	0.0145	0.1918	1	0
rs144040856	9	136521738	g	a	3e-04	-0.0756	0.135	0.5758	1	0
rs141021210	9	136521751	a	g	0	-0.131	0.997	0.8956	1	0
rs201973877	9	136522317	t	c	0.0003	-0.0251	0.0565	0.6566	1	0
rs75512464	9	136523487	a	t	0.002	-0.0264	0.0236	0.2621	1	0
rs76316834	9	136523555	g	a	4e-04	0.0484	0.13	0.709	1	0
rs3025380	9	136501756	g	c	0.0045	-0.0884	0.0119	1.096e-13	2	-0.082
rs74853476	9	136501834	t	c	0.0021	-0.0774	0.0181	1.973e-05	2	-0.082
rs142383279	9	136507332	g	a	0.0018	-0.0652	0.0194	0.0007811	2	-0.082
rs145059403	9	136507425	g	a	0.001	-0.102	0.0261	9.144e-05	2	-0.082
rs148439785	9	136521726	g	a	8e-04	-0.0595	0.0313	0.05706	2	-0.082
rs151228388	9	136522272	a	g	0.0013	-0.317	0.102	0.001909	2	-0.082

Supplementary Table 3.15 Clustering results of rare variants within the NPR1 gene with SBP

rsID	Chr	Position	Major allele	Minor allele	MAF	Effect	SE	P-value	Cluster membership	Cluster mean
rs56019647	1	153652129	c	t	1e-04	-0.113	0.165	0.4947	1	0
rs28730726	1	153653757	g	c	0.0016	-0.0175	0.0432	0.6842	1	0
rs199612927	1	153654234	g	a	2e-04	0.0688	0.0871	0.4298	1	0
rs140425746	1	153655966	g	a	2e-04	-0.0333	0.154	0.8287	1	0
rs201746049	1	153656216	a	g	0.0003	-0.033	0.195	0.8658	1	0
rs115938602	1	153656228	c	a	0	-0.139	0.29	0.6323	1	0
rs149202797	1	153658306	c	a	2e-04	-0.0203	0.136	0.8808	1	0
rs116775696	1	153659550	a	g	0	-0.333	0.458	0.4672	1	0
rs139174442	1	153660154	c	g	10.0e-05	0.297	0.35	0.3973	1	0
rs201787421	1	153660625	g	a	2e-04	-0.68	0.316	0.0312	1	0
rs61757359	1	153658297	g	a	0.0034	-0.0819	0.014	4.487e-09	2	-0.0846
rs61758562	1	153659131	g	a	5e-04	-0.133	0.0495	0.007185	2	-0.0846
rs116245325	1	153665650	c	t	8e-04	0.166	0.0287	7.686e-09	3	0.164

Supplementary Table 3.16 Clustering results of rare variants within the PLCB3 gene with SBP

rsID	Chr	Position	Major allele	Minor allele	MAF	Effect	SE	P-value	Cluster membership	Cluster mean
rs200996360	11	640253	c	g	0.0003	-0.03	0.0527	0.5689	1	0
rs111961110	11	64021930	a	g	0.0003	-0.14	0.207	0.4974	1	0
rs144191345	11	64022785	c	t	2e-04	-0.0212	0.108	0.8442	1	0
rs199645363	11	64022873	c	t	3e-04	0.0912	0.101	0.3668	1	0
rs138400940	11	64022905	c	t	1e-04	-0.381	0.576	0.5084	1	0
rs188572550	11	64026357	g	a	0	0.801	0.704	0.2551	1	0
rs201842672	11	64027567	g	t	0.0038	-0.0627	0.0711	0.3777	1	0
rs200923408	11	64028916	g	a	3e-04	-0.538	0.235	0.02196	1	0
rs200263631	11	64031571	t	c	0.0003	0.0847	0.0703	0.2283	1	0
rs79573066	11	64032784	g	a	0.0036	0.0262	0.0696	0.7066	1	0
rs61757725	11	64033360	g	t	0.009	0.0786	0.0599	0.1896	1	0
rs148059922	11	64033391	g	c	4e-04	-0.0149	0.131	0.9093	1	0
rs201342752	11	64033986	c	t	0.002	-0.0891	0.0706	0.2068	1	0
rs141163685	11	64034734	g	t	0.0031	0.249	0.113	0.02777	1	0
rs113554478	11	64034975	g	a	0.0011	0.00854	0.102	0.9333	1	0
rs117874826	11	64027666	a	c	0.0138	0.0467	0.00729	1.535e-10	2	0.0529
rs145502455	11	64031030	g	a	0.0054	0.0723	0.0118	8.563e-10	2	0.0529
rs142330950	11	64032945	g	t	0.0021	0.0456	0.0182	0.01236	2	0.0529

Supplementary Table 3.17 Clustering results of rare variants within the DBH gene with DBP

rsID	Chr	Position	Major allele	Minor allele	MAF	Effect	SE	P-value	Cluster membership	Cluster mean
rs76856960	9	136501569	g	a	0.0046	-0.00876	0.0138	0.5248	1	0
rs143544421	9	136501599	g	a	5e-04	0.0145	0.156	0.9262	1	0
rs78445536	9	136501746	g	a	5e-04	-0.0651	0.0572	0.255	1	0
rs146922432	9	136501768	g	a	3e-04	-0.125	0.144	0.3888	1	0
rs76819676	9	136507375	g	a	2e-04	0.0357	0.111	0.7477	1	0
rs200430427	9	136507456	c	t	4e-04	0.0385	0.154	0.8029	1	0
rs143535251	9	136507474	c	t	0.0013	0.00881	0.0248	0.7226	1	0
rs200628504	9	136507528	c	t	4e-04	-0.081	0.128	0.5253	1	0
rs5321	9	136507559	g	c	5e-04	0.0583	0.069	0.3984	1	0
rs199734841	9	136507586	g	c	4e-04	-0.225	0.2	0.2601	1	0
rs13306301	9	136508640	g	a	1e-04	-0.753	0.283	0.007722	1	0
rs5324	9	136508658	g	a	6e-04	-0.0431	0.0385	0.2625	1	0
rs145655199	9	136508682	g	a	0.001	0.129	0.18	0.473	1	0
rs201681337	9	136508691	g	a	0.002	0.0397	0.0755	0.5993	1	0
rs75215331	9	136513028	c	t	0.0034	0.00505	0.0168	0.7633	1	0
rs41316996	9	136521654	g	a	0.0032	0.00855	0.0146	0.557	1	0
rs148439785	9	136521726	g	a	8e-04	-0.0323	0.0313	0.3019	1	0
rs144040856	9	136521738	g	a	3e-04	-0.217	0.135	0.1089	1	0
rs141021210	9	136521751	a	g	0	-1.023	0.995	0.3037	1	0
rs201973877	9	136522317	t	c	0.0003	-0.0286	0.0566	0.6134	1	0
rs75512464	9	136523487	a	t	0.002	-0.025	0.0236	0.2893	1	0
rs76316834	9	136523555	g	a	4e-04	0.015	0.13	0.9078	1	0
rs3025380	9	136501756	g	c	0.0045	-0.103	0.0119	4.287e-18	2	-0.0927
rs74853476	9	136501834	t	c	0.0021	-0.0954	0.0182	1.529e-07	2	-0.0927
rs142383279	9	136507332	g	a	0.0018	-0.0701	0.0194	0.0003159	2	-0.0927
rs145059403	9	136507425	g	a	0.001	-0.0829	0.0262	0.001551	2	-0.0927
rs151228388	9	136522272	a	g	0.0013	-0.226	0.102	0.02711	2	-0.0927

Supplementary Table 3.18 Clustering results of rare variants within the PLCB3 gene with DBP

rsID	Chr	Position	Major allele	Minor allele	MAF	Effect	SE	P-value	Cluster membership	Cluster mean
rs200996360	11	640253	c	g	0.0003	0.0237	0.0528	0.653	1	0
rs111961110	11	64021930	a	g	0.0003	0.00639	0.207	0.9754	1	0
rs144191345	11	64022785	c	t	2e-04	0.0759	0.108	0.4822	1	0
rs199645363	11	64022873	c	t	3e-04	0.0863	0.101	0.3935	1	0
rs138400940	11	64022905	c	t	1e-04	-0.858	0.576	0.1362	1	0
rs188572550	11	64026357	g	a	0	0.624	0.704	0.3759	1	0
rs201842672	11	64027567	g	t	0.0038	-0.0781	0.0711	0.2721	1	0
rs200923408	11	64028916	g	a	3e-04	-0.194	0.235	0.4077	1	0
rs79573066	11	64032784	g	a	0.0036	0.0198	0.0697	0.7768	1	0
rs61757725	11	64033360	g	t	0.009	0.0537	0.06	0.3709	1	0
rs148059922	11	64033391	g	c	4e-04	-0.154	0.131	0.2409	1	0
rs201342752	11	64033986	c	t	0.002	-0.0181	0.0706	0.7979	1	0
rs141163685	11	64034734	g	t	0.0031	0.214	0.113	0.05838	1	0
rs113554478	11	64034975	g	a	0.0011	0.0951	0.102	0.3518	1	0
rs117874826	11	64027666	a	c	0.0138	0.041	0.00731	2.069e-08	2	0.0473
rs145502455	11	64031030	g	a	0.0054	0.0668	0.0118	1.619e-08	2	0.0473
rs200263631	11	64031571	t	c	0.0003	0.121	0.0704	0.08613	2	0.0473
rs142330950	11	64032945	g	t	0.0021	0.0367	0.0182	0.04445	2	0.0473

Supplementary Table 3.19 Clustering results of rare variants within the CEP120 gene with PP

rsID	Chr	Position	Major allele	Minor allele	MAF	Effect	SE	P-value	Cluster membership	Cluster mean
rs145436175	5	122682248	t	c	0.0008	0.233	0.268	0.385	1	0
rs140306974	5	122685717	g	a	0.0011	0.0594	0.126	0.6367	1	0
rs200061679	5	122685731	t	a	0.0015	0.0692	0.0388	0.07491	1	0
rs142792779	5	122708381	c	t	3e-04	-0.0136	0.213	0.9492	1	0
rs139865050	5	122713092	c	g	0	-0.192	0.707	0.7861	1	0
rs74938108	5	122713159	c	t	0.0016	-0.0257	0.0905	0.7765	1	0
rs201600892	5	122713191	c	g	10.0e-05	0.0767	0.447	0.8637	1	0
rs61744334	5	122714044	t	c	0.0006	-0.0201	0.0684	0.7689	1	0
rs144490830	5	122714104	g	c	2e-04	0.37	0.179	0.03908	1	0
rs147277049	5	122720724	t	c	0.0006	-0.0181	0.036	0.6152	1	0
rs200450605	5	122725693	g	a	0.0017	0.016	0.0229	0.4841	1	0
rs201571160	5	122725754	c	g	0.0003	0.166	0.316	0.6004	1	0
rs114281792	5	122725761	t	c	0.0067	-0.00233	0.0098	0.8124	1	0
rs61747983	5	122725768	g	a	1e-04	0.0156	0.112	0.8895	1	0
rs201955087	5	122727015	c	t	0.0013	0.274	0.106	0.009927	1	0
rs147273517	5	122748194	t	c	10.0e-05	0.36	0.162	0.02599	1	0
rs202103949	5	122748198	t	c	0.0004	-0.029	0.121	0.8106	1	0
rs199793672	5	122758665	t	c	10.0e-05	-0.479	0.503	0.3404	1	0
rs2303720	5	122682334	c	t	0.0291	-0.0418	0.00482	4.532e-18	2	-0.0435
rs114280473	5	122714092	g	a	0.0063	-0.0632	0.011	8.076e-09	2	-0.0435
rs189429890	5	122729025	c	t	0.0065	-0.0373	0.00999	0.0001886	2	-0.0435

Supplementary Table 3.20 Clustering results of rare variants within COL21A1 gene with PP

rsID	Chr	Position	Major allele	Minor allele	MAF	Effect	SE	P-value	Cluster membership	Cluster mean
rs200478915	6	55923968	g	c	8e-04	-0.0449	0.0323	0.1652	1	0
rs200564236	6	55925008	g	a	8e-04	-0.0212	0.162	0.8957	1	0
rs199722485	6	55925588	t	g	0.0005	-0.0128	0.106	0.9038	1	0
rs200674177	6	55925689	g	a	2e-04	-0.0981	0.189	0.6046	1	0
rs201839603	6	55925762	a	c	10.0e-05	0.0193	0.577	0.9733	1	0
rs9464337	6	55925801	g	t	0.0054	0.000204	0.0116	0.986	1	0
rs201892311	6	55926464	g	c	1e-04	0.426	0.707	0.547	1	0
rs199910287	6	55935556	g	a	6e-04	0.054	0.0445	0.2255	1	0
rs191626317	6	55939059	g	t	0.002	0.0424	0.102	0.6783	1	0
rs75605879	6	55966311	a	g	10.0e-05	-0.147	0.107	0.169	1	0
rs202115077	6	55988871	g	t	0.002	0.00866	0.0186	0.6409	1	0
rs201267383	6	56006611	t	c	0.0003	-0.077	0.0817	0.3461	1	0
rs35583895	6	56006732	a	g	0.0081	0.00138	0.0479	0.9771	1	0
rs200361985	6	56029264	g	a	0.002	0.0616	0.0546	0.2593	1	0
rs142653960	6	56035494	c	t	4e-04	-0.0656	0.0525	0.2115	1	0
rs199532612	6	56035853	a	t	0.0017	-0.0186	0.0217	0.3925	1	0
rs200708113	6	56035881	c	t	5e-04	-0.0112	0.0509	0.8259	1	0
rs202026963	6	56035909	g	a	1e-04	0.0193	0.242	0.9364	1	0
rs147394600	6	56047400	g	a	1e-04	-0.159	0.164	0.331	1	0
rs200999181	6	55935568	c	a	0.0012	0.336	0.0244	3.471e-43	2	0.335
rs115079907	6	55924005	c	t	0.0015	0.207	0.0248	5.573e-17	3	0.171
rs76146749	6	55925783	t	a	7e-04	0.198	0.0782	0.01144	3	0.171
rs200401514	6	55989091	c	t	4e-04	0.124	0.0552	0.02532	3	0.171
rs2764043	6	56035643	a	g	0.0016	0.152	0.0203	7.82e-14	3	0.171

Supplementary Table 3.21 Clustering results of rare variants within the NOX4 gene with PP

rsID	Chr	Position	Major allele	Minor allele	MAF	Effect	SE	P-value	Cluster membership	Cluster mean
rs115031759	11	89073269	g	a	1e-04	0.192	0.161	0.2323	1	0
rs201165492	11	89106599	c	t	1e-04	-0.274	0.576	0.6345	1	0
rs149515506	11	89135492	a	g	10.0e-05	0.428	0.706	0.5445	1	0
rs147350656	11	89177310	c	t	1e-04	-0.184	0.236	0.4375	1	0
rs142433357	11	89182609	t	c	10.0e-05	-1.231	0.707	0.08181	1	0
rs55977241	11	89182652	c	t	0.002	0.000803	0.109	0.9941	1	0
rs145686545	11	89224387	c	t	2e-04	0.162	0.202	0.4229	1	0
rs144215891	11	89069094	t	c	0.0014	-0.0716	0.0302	0.01798	2	-0.0962
rs139341533	11	89182666	c	a	0.0043	-0.0905	0.0126	8.085e-13	2	-0.0962
rs56061986	11	89182686	t	c	0.0029	-0.113	0.0161	2.385e-12	2	-0.0962

Supplementary Table 3.22 Clustering results of rare variants within the DBH gene with HTN

rsID	Chr	Position	Major allele	Minor allele	MAF	Z value	P-value	Cluster membership	Cluster mean
rs76856960	9	136501569	g	a	0.0035	-0.511	0.6096	1	0
rs143544421	9	136501599	g	a	2e-04	-0.526	0.5989	1	0
rs78445536	9	136501746	g	a	3e-04	0.234	0.815	1	0
rs146922432	9	136501768	g	a	1e-04	1.257	0.2089	1	0
rs76819676	9	136507375	g	a	1e-04	-0.787	0.4313	1	0
rs143535251	9	136507474	c	t	0.0014	0.397	0.6916	1	0
rs200628504	9	136507528	c	t	2e-04	-0.0675	0.9462	1	0
rs5321	9	136507559	g	c	2e-04	0.226	0.8212	1	0
rs199734841	9	136507586	g	c	2e-04	0.897	0.3697	1	0
rs13306301	9	136508640	g	a	0	-1.378	0.1683	1	0
rs5324	9	136508658	g	a	8e-04	-1.404	0.1604	1	0
rs145655199	9	136508682	g	a	3e-04	-1.247	0.2125	1	0
rs201681337	9	136508691	g	a	2e-04	1.109	0.2676	1	0
rs75215331	9	136513028	c	t	0.0025	-0.0818	0.9348	1	0
rs41316996	9	136521654	g	a	0.003	2.028	0.04261	1	0
rs144040856	9	136521738	g	a	1e-04	0.0598	0.9523	1	0
rs141021210	9	136521751	a	g	0	-0.665	0.5059	1	0
rs201973877	9	136522317	t	c	0.0002	-0.923	0.356	1	0
rs75512464	9	136523487	a	t	0.0014	-0.289	0.7723	1	0
rs148806316	9	136523534	c	t	0.0025	0.295	0.7679	1	0
rs76316834	9	136523555	g	a	2e-04	-0.39	0.6962	1	0
rs77273740	9	136501728	c	t	0.0042	-2.933	0.003353	2	-2.612
rs3025380	9	136501756	g	c	0.0045	-5.297	1.176e-07	2	-2.612
rs74853476	9	136501834	t	c	0.0022	-2.186	0.02885	2	-2.612
rs142383279	9	136507332	g	a	0.002	-1.826	0.06781	2	-2.612
rs145059403	9	136507425	g	a	0.0012	-3.468	0.0005248	2	-2.612
rs200430427	9	136507456	c	t	1e-04	-2.306	0.02112	2	-2.612
rs148439785	9	136521726	g	a	0.0019	-2.196	0.02806	2	-2.612
rs151228388	9	136522272	a	g	0.0003	-2.399	0.01646	2	-2.612

Supplementary Table 3.23 Clustering results of rare variants within the NPR1 gene with HTN

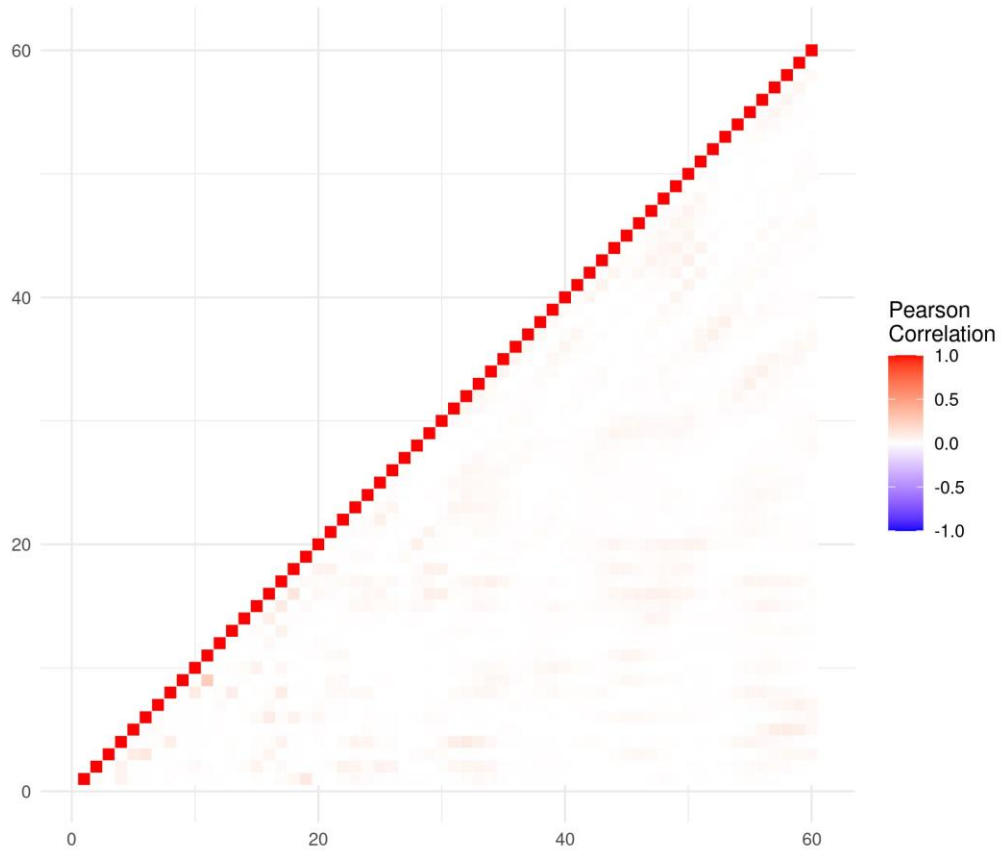
rsID	Chr	Position	Major allele	Minor allele	MAF	Z value	P-value	Cluster membership	Cluster mean
rs56019647	1	153652129	c	t	0	-0.727	0.4675	1	0
rs199612927	1	153654234	g	a	1e-04	0.468	0.6397	1	0
rs201746049	1	153656216	a	g	10.0e-05	-1.254	0.21	1	0
rs115938602	1	153656228	c	a	0	-1.174	0.2405	1	0
rs149202797	1	153658306	c	a	1e-04	0.23	0.8182	1	0
rs116775696	1	153659550	a	g	0	0.641	0.5213	1	0
rs139174442	1	153660154	c	g	0	0.0493	0.9606	1	0
rs201787421	1	153660625	g	a	1e-04	-0.287	0.7739	1	0
rs140425746	1	153655966	g	a	1e-04	-1.607	0.1081	2	-2.053
rs61757359	1	153658297	g	a	0.0036	-3.719	2e-04	2	-2.053
rs61758562	1	153659131	g	a	0.0011	-1.693	0.09054	2	-2.053
rs116245325	1	153665650	c	t	8e-04	4.335	1.457e-05	3	4.334

Supplementary Table 3.24 Clustering results of rare variants within the PLCB3 gene with HTN

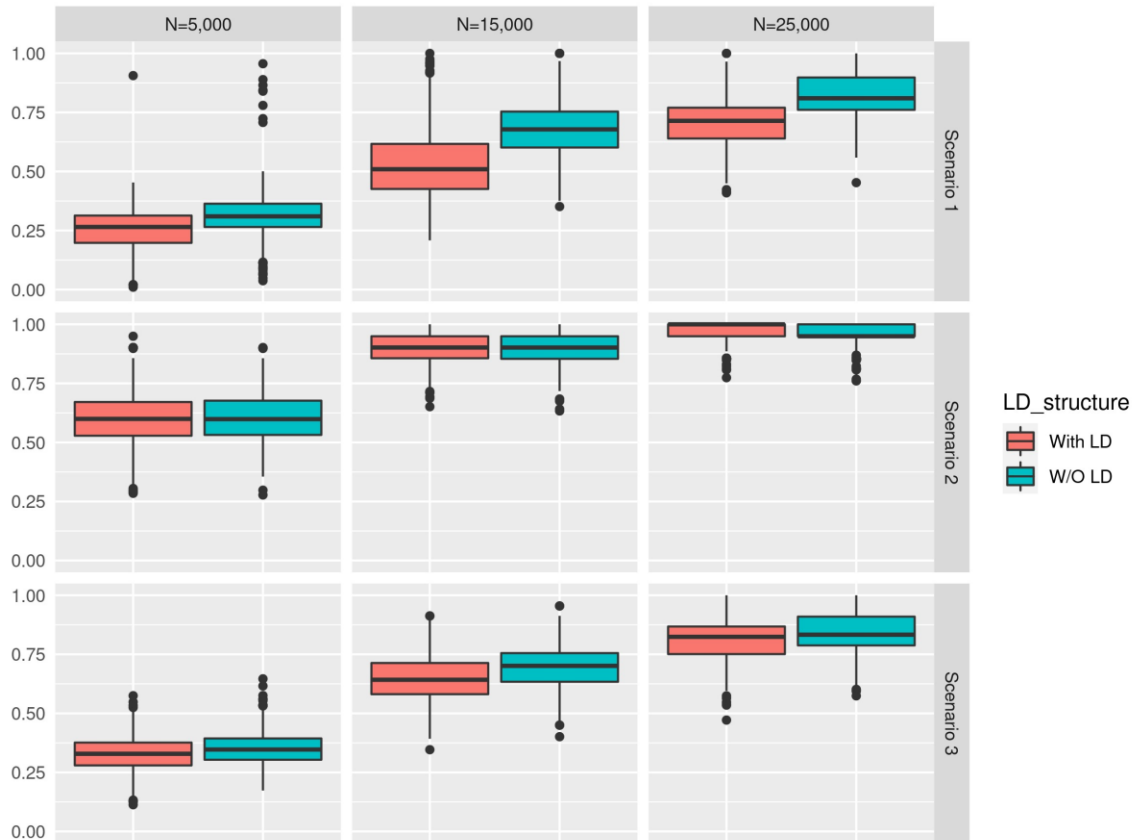
rsID	Chr	Position	Major allele	Minor allele	MAF	Z value	P-value	Cluster membership	Cluster mean
rs200996360	11	640253	c	g	0.0003	-0.905	0.3653	1	0
rs111961110	11	64021930	a	g	10.0e-05	-0.917	0.3591	1	0
rs144191345	11	64022785	c	t	1e-04	0.259	0.7958	1	0
rs199645363	11	64022873	c	t	2e-04	-0.587	0.5575	1	0
rs138400940	11	64022905	c	t	1e-04	-1.28	0.2004	1	0
rs188572550	11	64026357	g	a	0	-0.324	0.7457	1	0
rs201842672	11	64027567	g	t	9e-04	-0.58	0.5616	1	0
rs200923408	11	64028916	g	a	1e-04	-0.496	0.6198	1	0
rs200263631	11	64031571	t	c	0.0002	1.033	0.3018	1	0
rs79573066	11	64032784	g	a	2e-04	0.665	0.5063	1	0
rs142330950	11	64032945	g	t	0.0021	2.175	0.02966	1	0
rs61757725	11	64033360	g	t	3e-04	0.427	0.6694	1	0
rs148059922	11	64033391	g	c	1e-04	1.308	0.1908	1	0
rs201342752	11	64033986	c	t	5e-04	1.223	0.2215	1	0
rs141163685	11	64034734	g	t	4e-04	0.867	0.3859	1	0
rs113554478	11	64034975	g	a	1e-04	0.708	0.4787	1	0
rs117874826	11	64027666	a	c	0.0308	5.712	1.119e-08	2	4.808
rs145502455	11	64031030	g	a	0.0047	4.015	5.952e-05	2	4.808

B2. Supplementary Figures

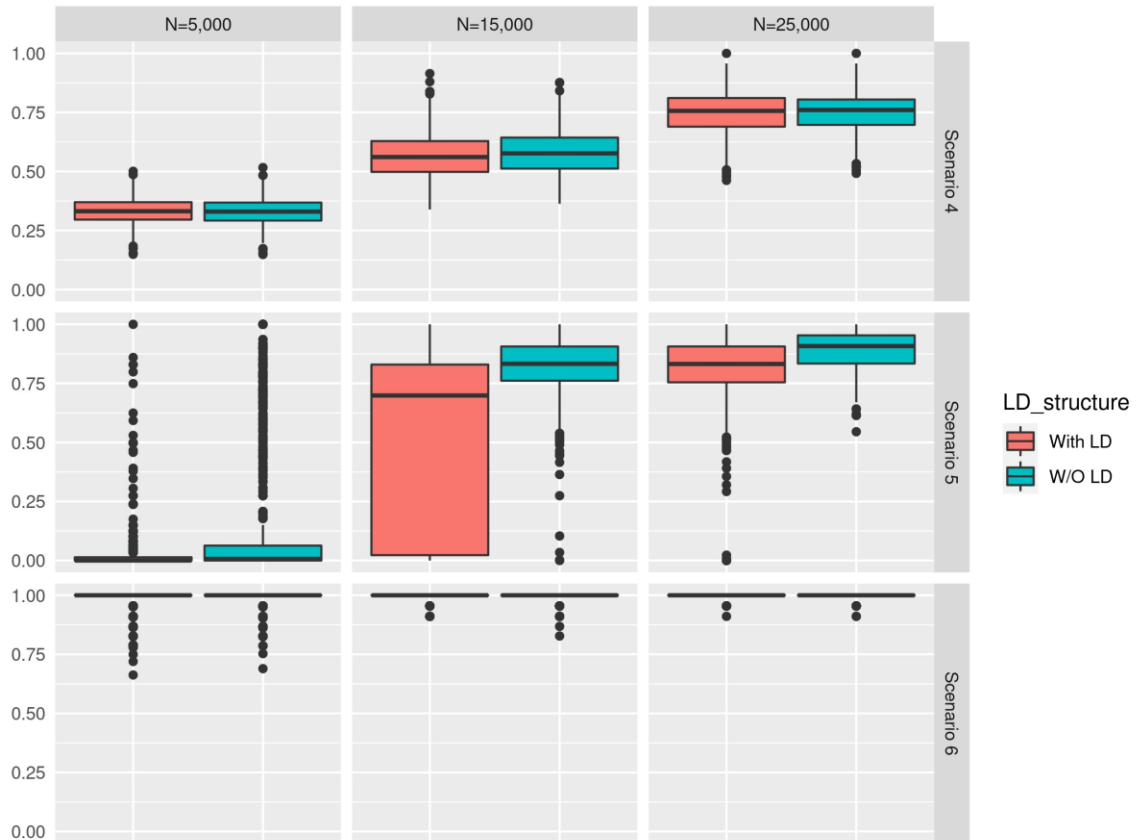
Supplementary Figure 3.1 Heatmap of the average correlation matrix of the 60 simulated rare variants with the presence of LD structure



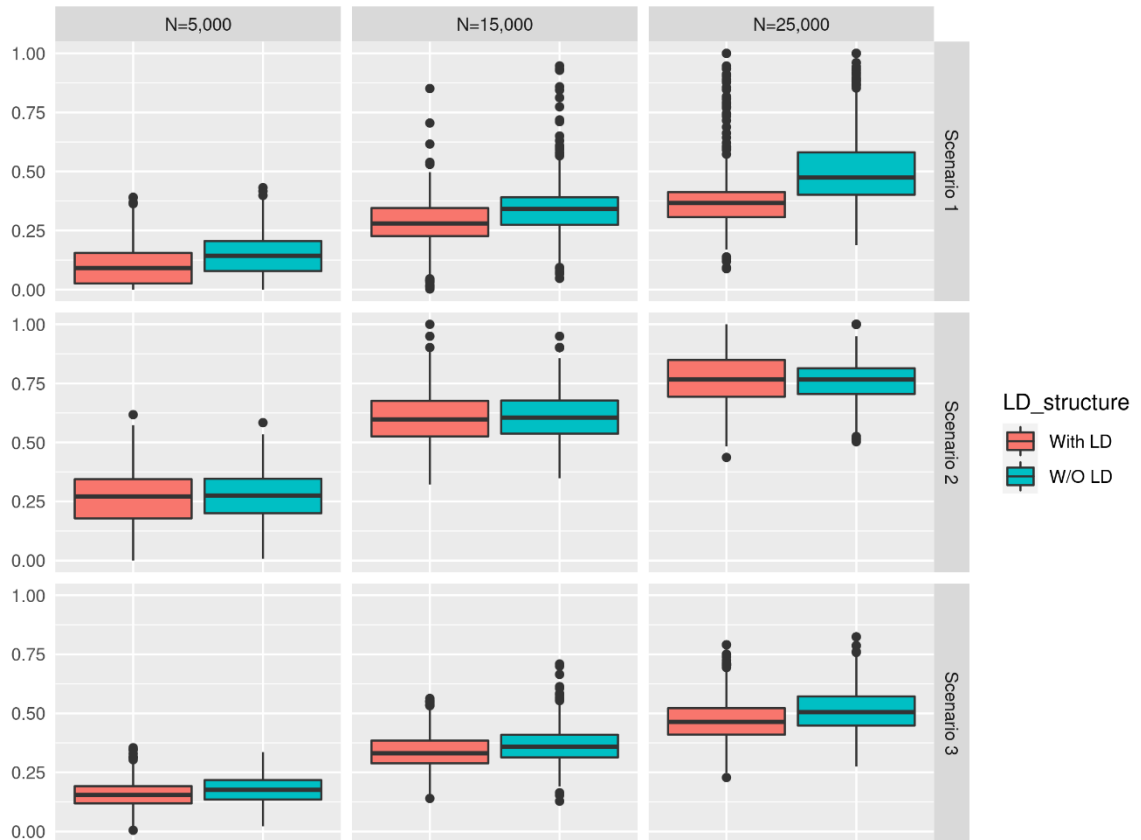
Supplementary Figure 3.2 Boxplots of ARI values of simulation scenarios 1, 2, and 3 with combined weighting scheme with/without LD for a continuous trait



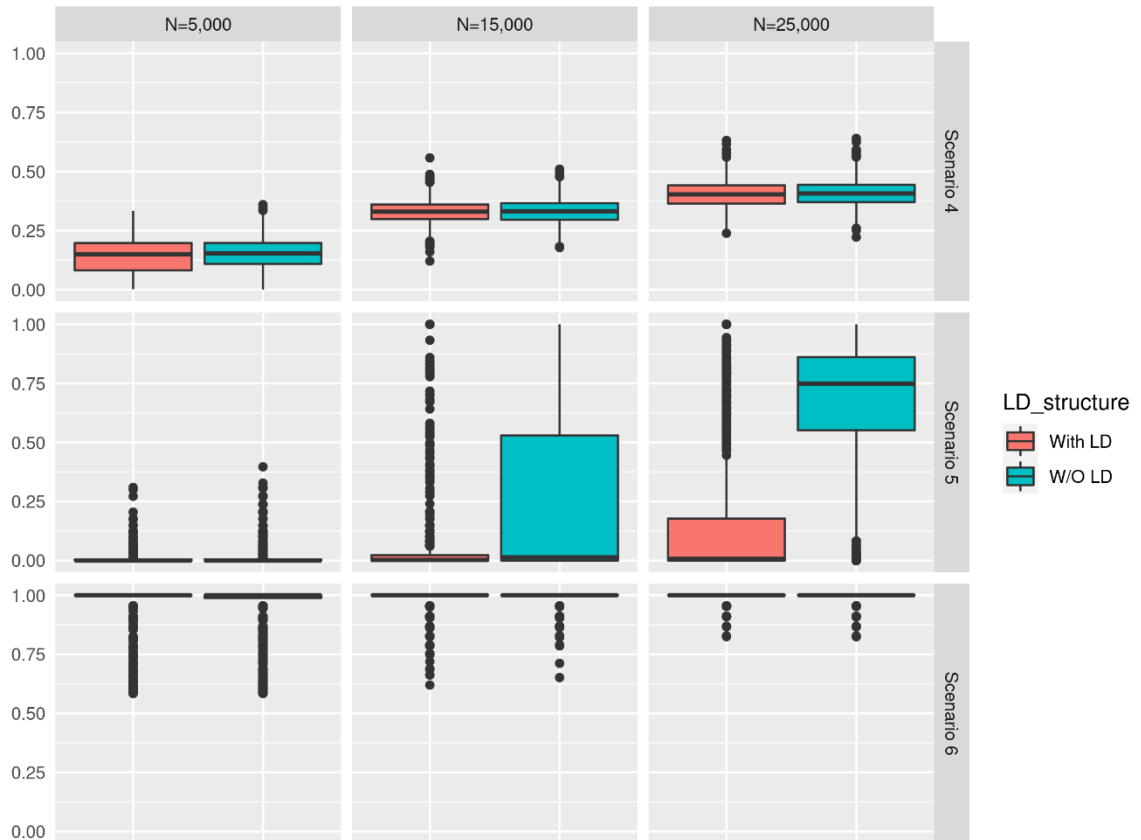
Supplementary Figure 3.3 Boxplots of ARI values of simulation scenarios 4, 5, and 6 with combined weighting scheme with/without LD for a continuous trait



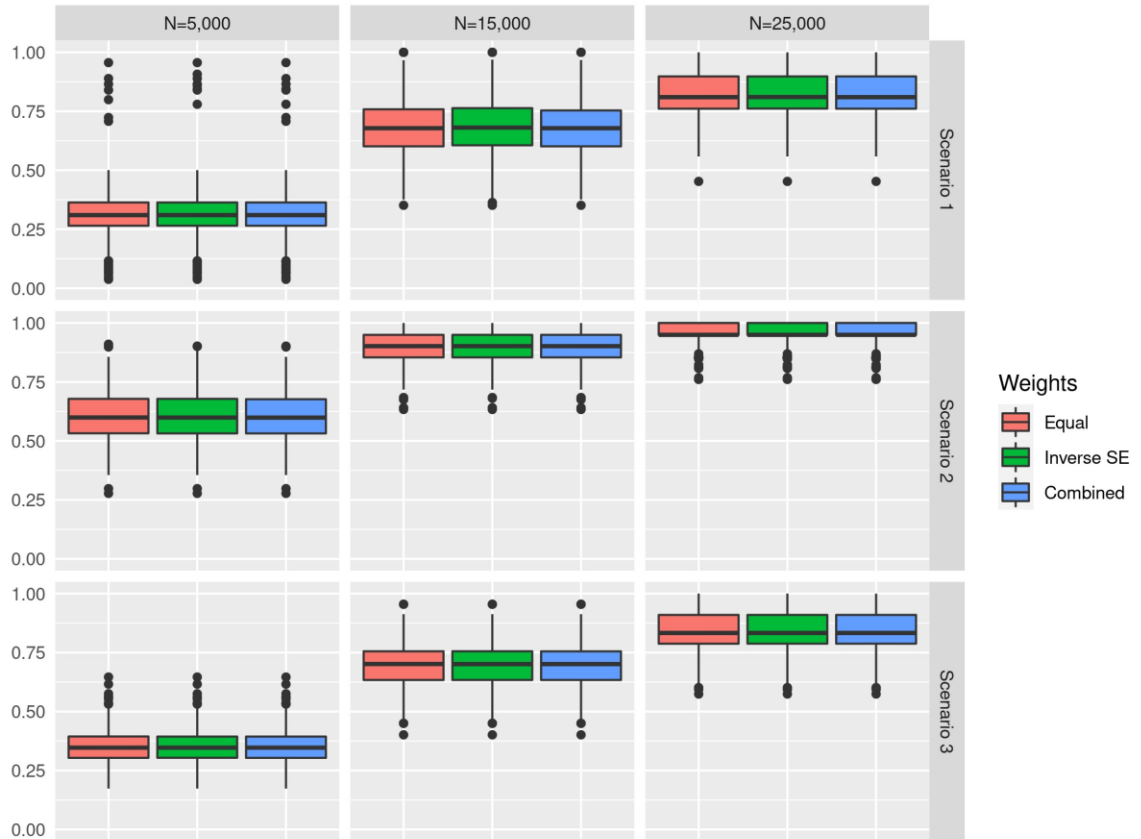
Supplementary Figure 3.4 Boxplots of ARI values of simulation scenarios 1, 2, and 3 with combined weighting scheme with/without LD for a binary trait



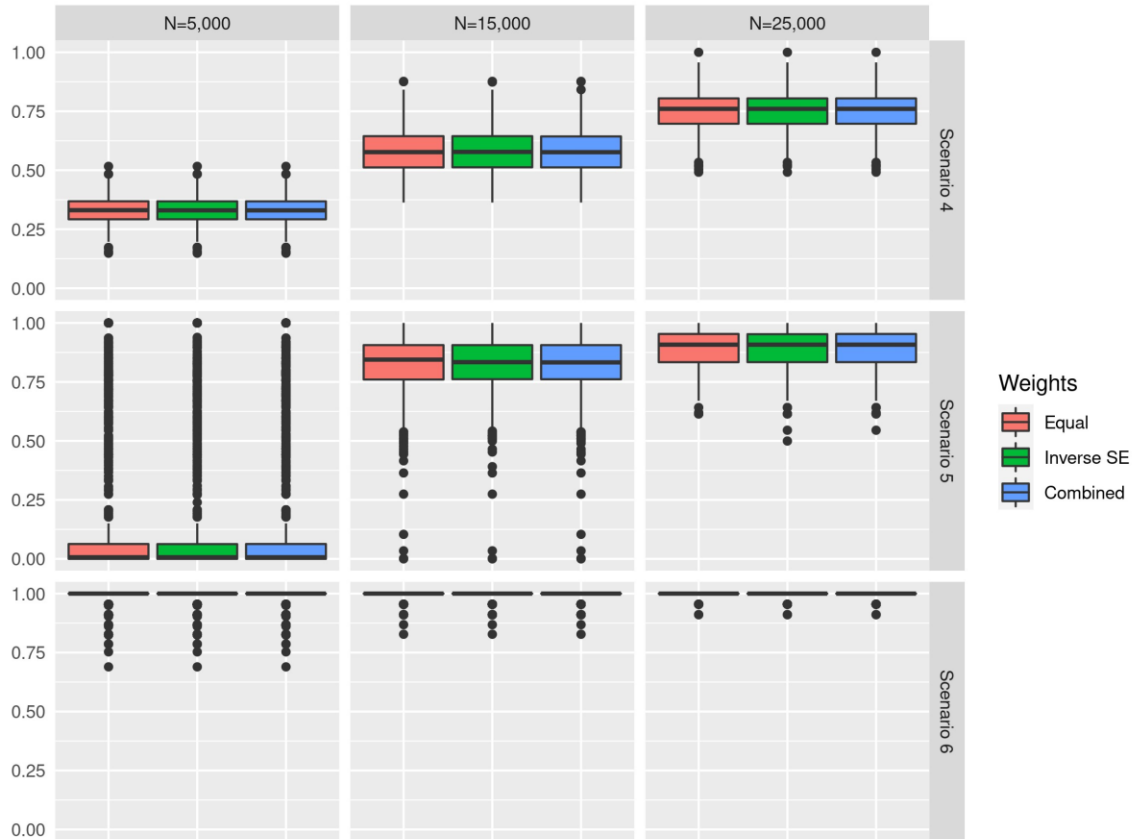
Supplementary Figure 3.5 Boxplots of ARI values of simulation scenarios 4, 5, and 6 with combined weighting scheme with/without LD for a binary trait



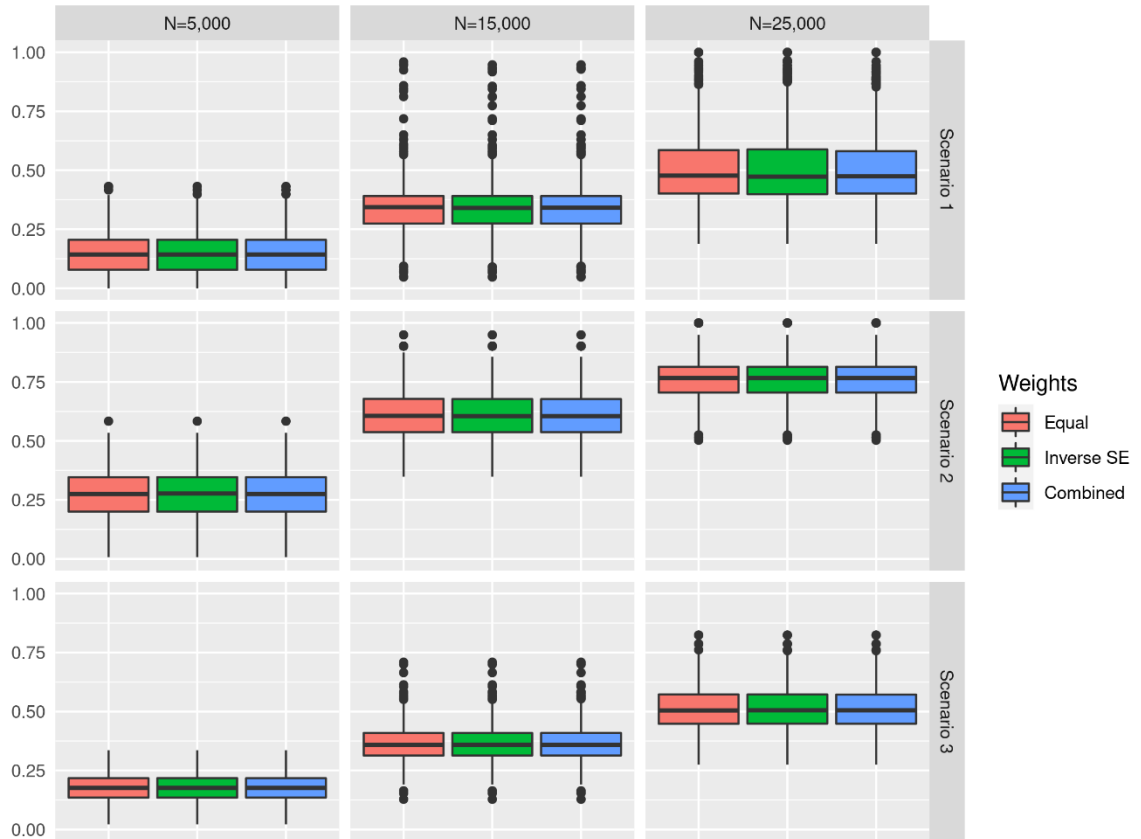
Supplementary Figure 3.6 Boxplots of ARI values of simulation scenarios 1, 2, and 3 with equal, inverse SE and combined weighting schemes with the absence of LD for a continuous trait



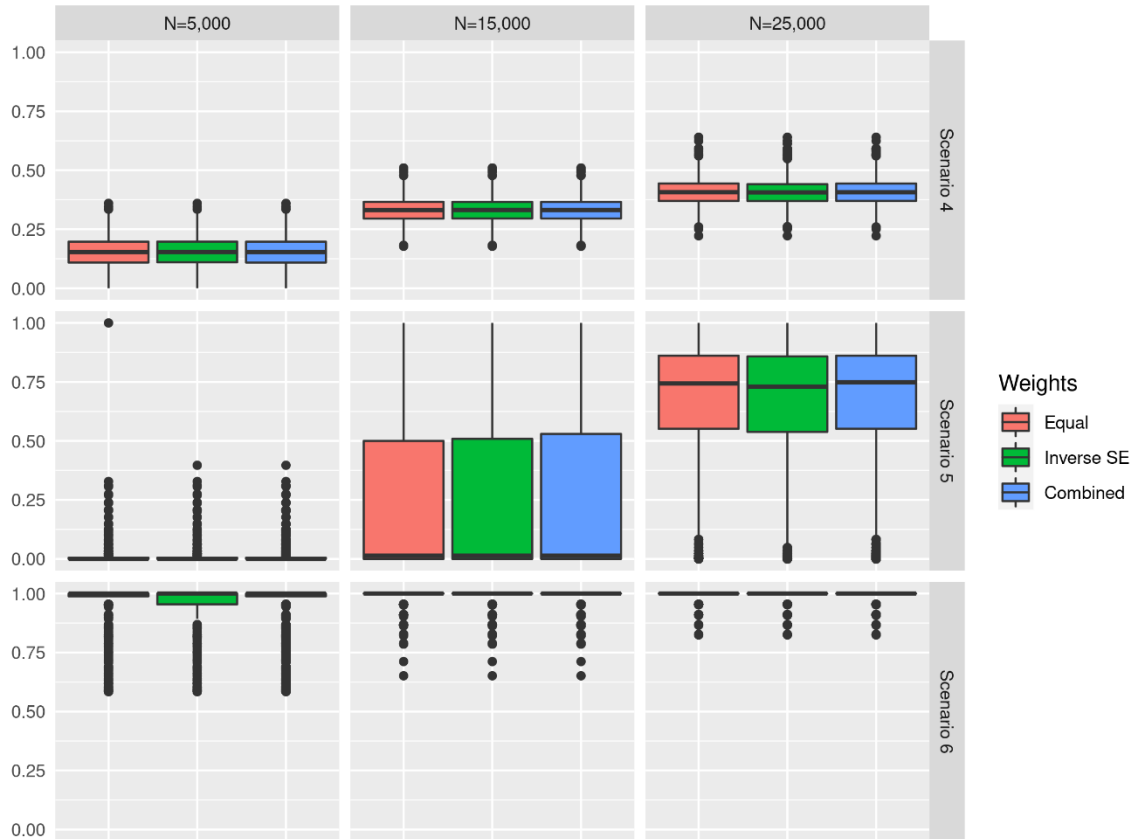
Supplementary Figure 3.7 Boxplots of ARI values of simulation scenarios 4, 5, and 6 with equal, inverse SE and combined weighting schemes with the absence of LD for a continuous trait



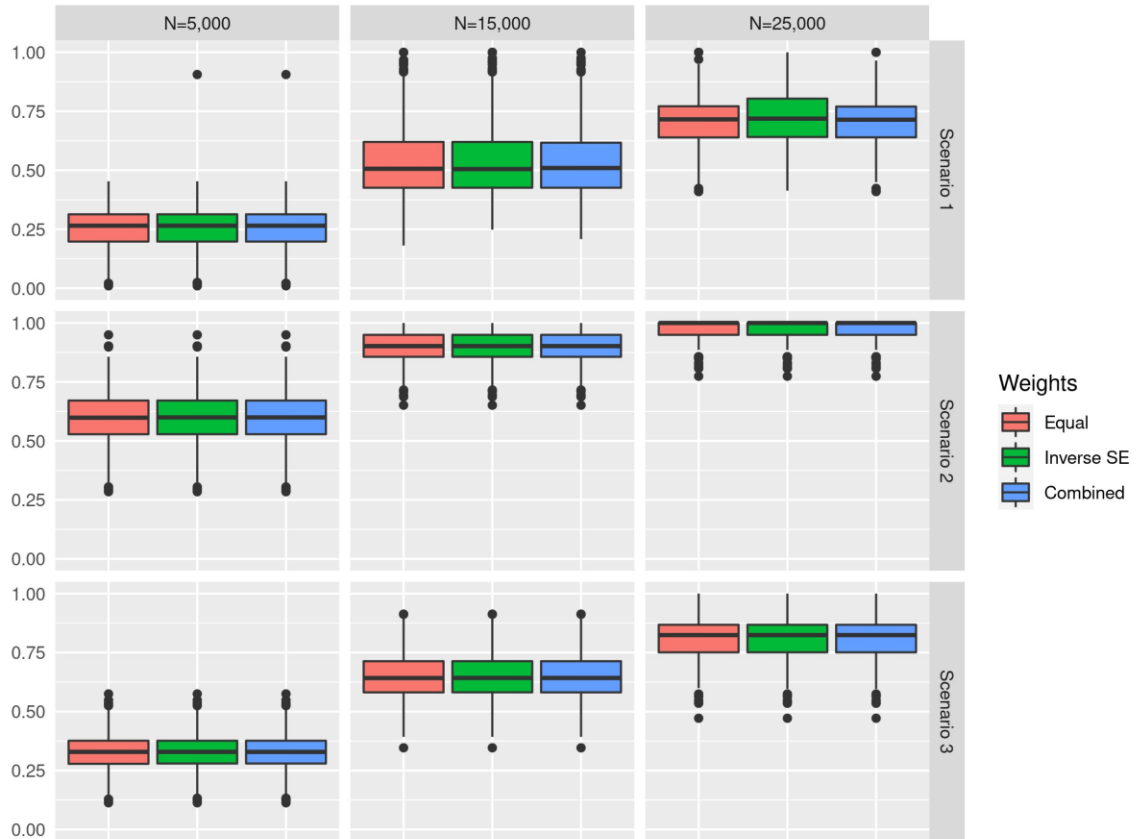
Supplementary Figure 3.8 Boxplots of ARI values of simulation scenarios 1, 2, and 3 with equal, inverse SE and combined weighting schemes with the absence of LD for a binary trait



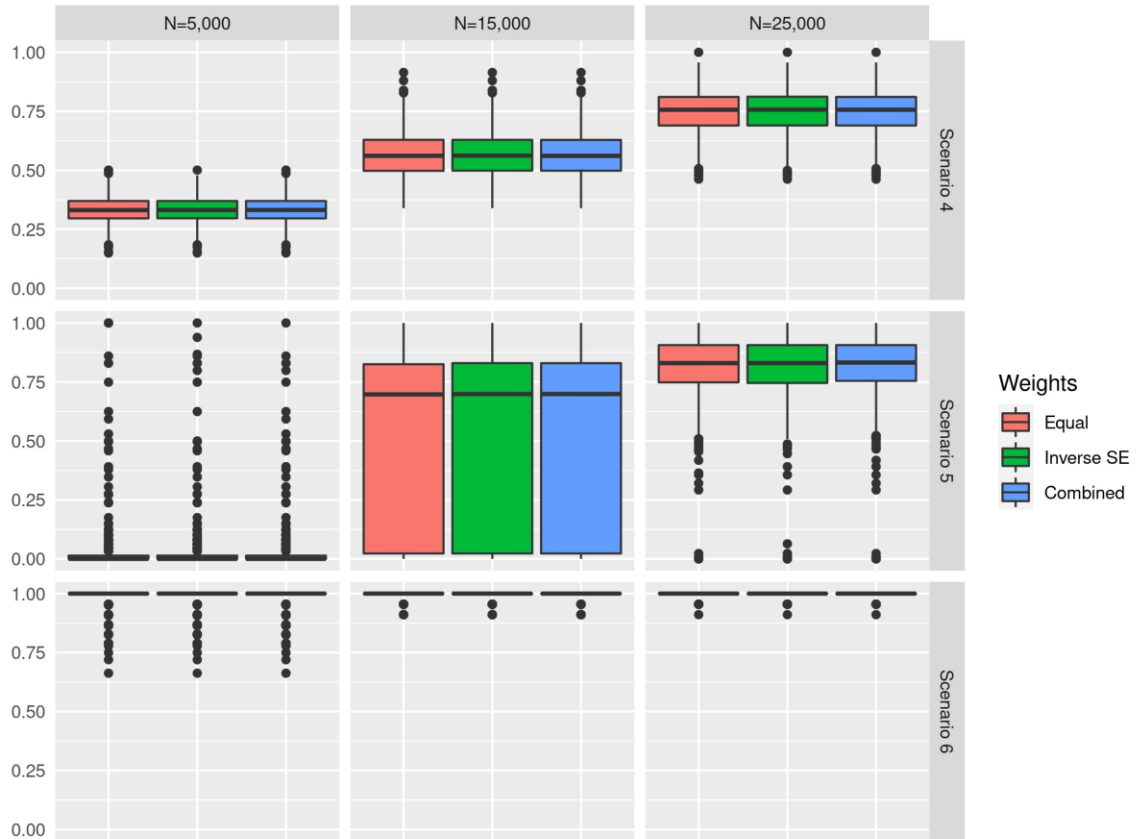
Supplementary Figure 3.9 Boxplots of ARI values of simulation scenarios 4, 5, and 6 with equal, inverse SE and combined weighting schemes with the absence of LD for a binary trait



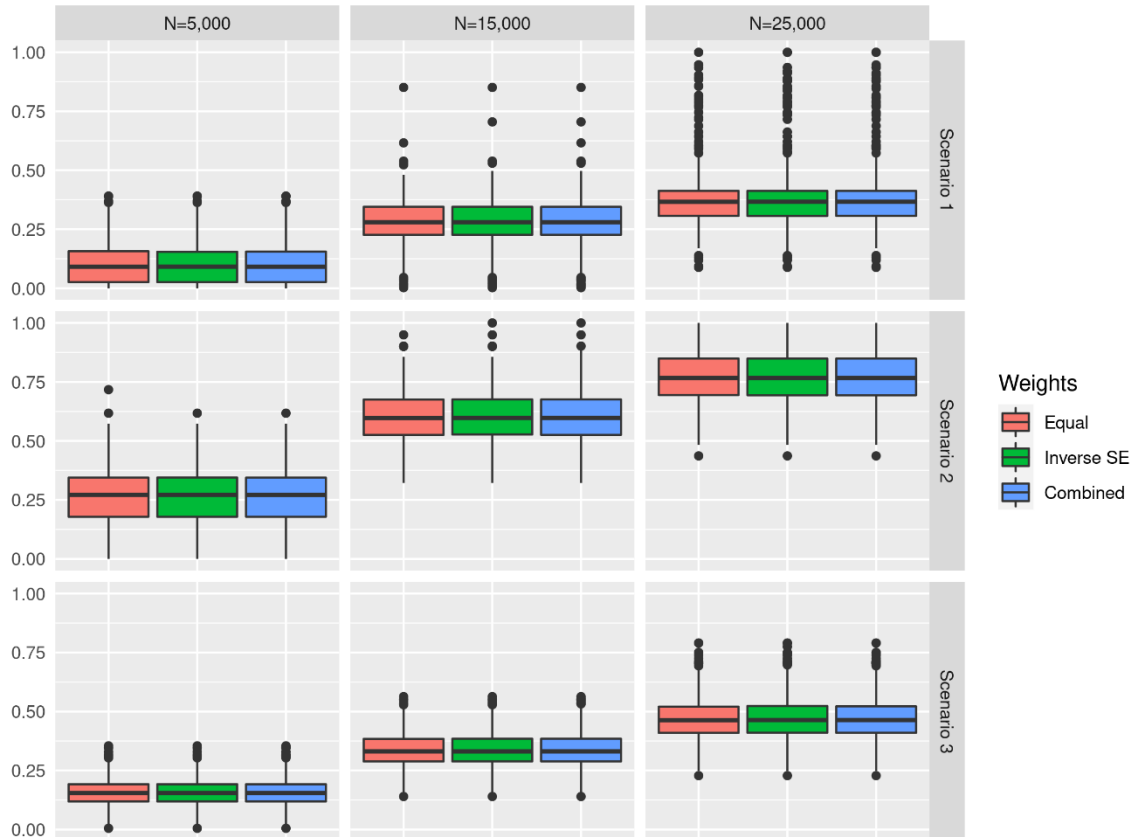
Supplementary Figure 3.10 Boxplots of ARI values of simulation scenarios 1, 2, and 3 with equal, inverse SE and combined weighting schemes with the presence of LD for a continuous trait



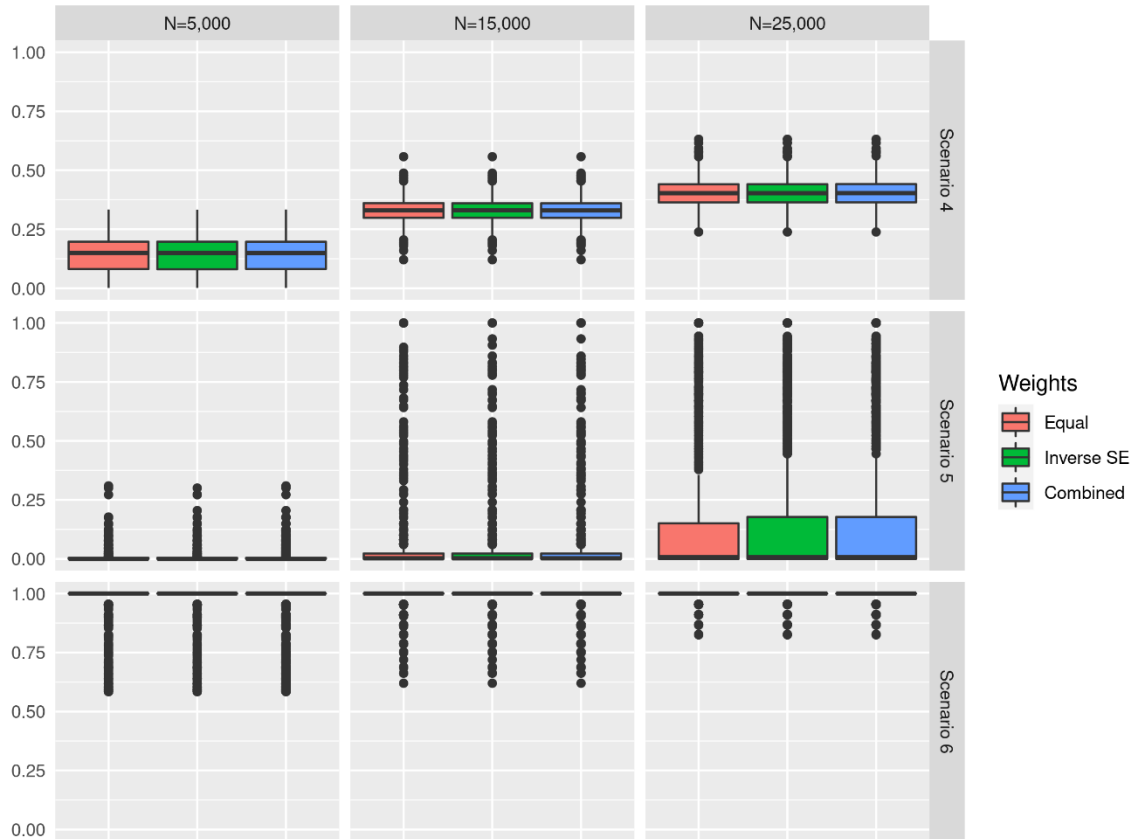
Supplementary Figure 3.11 Boxplots of ARI values of simulation scenarios 4, 5, and 6 with equal, inverse SE and combined weighting schemes with the presence of LD for a continuous trait



Supplementary Figure 3.12 Boxplots of ARI values of simulation scenarios 1, 2, and 3 with equal, inverse SE and combined weighting schemes with the presence of LD for a binary trait



Supplementary Figure 3.13 Boxplots of ARI values of simulation scenarios 4, 5, and 6 with equal, inverse SE and combined weighting schemes with the presence of LD for a binary trait



APPENDIX C: SUPPLEMENTARY MATERIALS FOR CHAPTER 4

Supplementary Table 4.1 Associations between heteroplasmic variants and BMI across sixteen mtDNA genes among AA

Gene	Definition 1						Definition 2						Definition 3					
	beta	se	burden	skat	skato	acat	beta	se	burden	skat	skato	acat	beta	se	burden	skat	skato	acat
D-loop	-0.2	0.18	0.26	0.083	0.14	0.13	-0.006	0.0066	0.36	0.0097	0.012	0.019	-1.78	2.79	0.52	0.2	0.34	0.32
MT-RNR1	-0.36	0.43	0.41	0.53	0.59	0.47	-0.0029	0.014	0.84	0.46	0.47	0.72	-1.03	0.87	0.24	0.22	0.36	0.23
MT-RNR2	0.029	0.33	0.93	0.38	0.57	0.85	0.00097	0.01	0.92	0.044	0.54	0.18	-0.29	0.79	0.71	0.64	0.86	0.68
MT-ND1	0.052	0.36	0.89	0.61	0.82	0.82	0.00011	0.011	0.99	0.57	0.66	0.98	-0.33	1.75	0.85	0.76	0.93	0.82
MT-ND2	0.43	0.51	0.4	0.063	0.12	0.12	0.006	0.014	0.66	0.45	0.68	0.57	-1.92	1.76	0.28	0.6	0.44	0.42
MT-CO1	-0.31	0.22	0.16	0.22	0.23	0.18	-0.013	0.0079	0.1	0.047	0.097	0.064	-0.58	0.97	0.55	0.15	0.26	0.27
MT-CO2	-0.92	0.56	0.097	0.45	0.16	0.17	-0.026	0.017	0.12	0.096	0.032	0.11	-4.46	2.01	0.026	0.66	0.047	0.055
MT-ATP8	1.25	1	0.21	0.34	0.34	0.27	0.029	0.028	0.31	0.42	0.59	0.36	-0.68	6.84	0.92	0.91	0.97	0.92
MT-ATP6	-0.26	0.43	0.55	0.56	0.76	0.56	-0.0029	0.013	0.83	0.65	0.79	0.76	0.26	6.87	0.97	0.65	0.86	0.94
MT-CO3	-0.51	0.53	0.33	0.15	0.26	0.21	-0.018	0.016	0.26	0.4	0.28	0.32	-0.9	1.84	0.62	0.68	0.84	0.65
MT-ND3	2.07	0.88	0.018	0.035	0.031	0.024	0.08	0.027	0.0026	0.0012	0.45	0.0017	0.29	2.21	0.89	0.28	0.43	0.76
MT-ND4L	-0.21	0.98	0.83	0.07	0.11	0.2	-0.0044	0.029	0.88	0.044	0.046	0.13	1.34	2.85	0.64	0.0044	0.0061	0.0089
MT-ND4	-0.11	0.35	0.75	0.56	0.77	0.67	-0.0031	0.011	0.78	0.2	0.55	0.48	1.67	1.34	0.21	0.07	0.12	0.11
MT-ND5	-0.21	0.22	0.35	0.33	0.5	0.34	-0.0043	0.0066	0.52	0.15	0.52	0.25	-1.03	1.08	0.34	0.89	0.52	0.77
MT-ND6	-0.25	0.52	0.63	0.33	0.51	0.47	-0.015	0.017	0.37	0.41	0.48	0.39	0.33	2.51	0.9	0.32	0.5	0.77
MT-CYB	0.44	0.37	0.23	0.039	0.073	0.068	0.016	0.01	0.12	0.0044	0.039	0.0086	-1.46	1.96	0.46	0.94	0.68	0.87
WG	-0.024	0.035	0.49	-	-	-	-0.00058	0.0011	0.61	-	-	-	-0.23	0.22	0.29	-	-	-

Supplementary Table 4.2 Associations between heteroplasmic variants and BMI across sixteen mtDNA genes among EA

Gene	Definition 1						Definition 2						Definition 3					
	beta	se	burden	skat	skato	acat	beta	se	burden	skat	skato	acat	beta	se	burden	skat	skato	acat
D-loop	-0.24	0.13	0.055	0.93	0.098	0.34	-0.0093	0.0042	0.025	0.97	0.13	0.34	-3.01	1.61	0.061	0.66	0.11	0.13
MT-RNR1	-0.2	0.26	0.44	0.31	0.52	0.37	-0.0036	0.0063	0.57	0.15	0.14	0.27	-0.11	0.48	0.82	0.28	0.47	0.62
MT-RNR2	0.07	0.21	0.74	0.91	0.92	0.86	0.0017	0.005	0.73	0.83	0.56	0.79	0.1	0.37	0.78	0.71	0.91	0.75
MT-ND1	-0.38	0.32	0.24	0.62	0.4	0.4	-0.0075	0.0069	0.28	0.64	0.48	0.44	-0.91	1	0.36	0.53	0.56	0.44
MT-ND2	-0.058	0.31	0.85	0.68	0.89	0.79	-0.0031	0.0068	0.65	0.85	0.84	0.79	1.34	0.98	0.17	0.14	0.25	0.16
MT-CO1	0.063	0.22	0.78	0.48	0.72	0.66	0.003	0.005	0.55	0.47	0.17	0.51	0.49	0.51	0.34	0.5	0.54	0.41
MT-CO2	0.42	0.33	0.21	0.16	0.27	0.18	0.01	0.0075	0.18	0.074	0.045	0.11	0.16	0.92	0.86	0.55	0.78	0.77
MT-ATP8	1.29	0.62	0.037	0.08	0.065	0.051	0.031	0.013	0.017	0.028	0.0057	0.021	10.96	6.96	0.12	0.43	0.17	0.19
MT-ATP6	-0.0037	0.31	0.99	0.044	0.083	0.98	0.00067	0.0068	0.92	0.073	0.28	0.45	-0.43	4.06	0.92	0.16	0.29	0.74
MT-CO3	0.07	0.32	0.83	0.87	0.97	0.85	0.0011	0.007	0.87	0.83	0.97	0.85	1.02	1.02	0.32	0.32	0.51	0.32
MT-ND3	0.015	0.51	0.98	0.45	0.69	0.95	0.0047	0.012	0.68	0.45	0.3	0.58	-0.96	1.26	0.45	0.56	0.64	0.5
MT-ND4L	-0.52	0.62	0.4	0.12	0.22	0.2	-0.012	0.013	0.37	0.14	0.43	0.21	-1.41	1.27	0.27	0.039	0.07	0.069
MT-ND4	-0.34	0.24	0.16	0.87	0.25	0.57	-0.0091	0.0059	0.12	0.92	0.43	0.69	-0.42	0.69	0.54	0.45	0.68	0.5
MT-ND5	-0.22	0.18	0.22	0.93	0.36	0.81	-0.0059	0.0044	0.18	0.82	0.099	0.5	-0.41	0.61	0.5	0.86	0.73	0.76
MT-ND6	0.4	0.36	0.27	0.6	0.44	0.41	0.005	0.008	0.53	0.39	0.73	0.46	3.03	1.38	0.028	0.083	0.051	0.042
MT-CYB	0.039	0.19	0.84	0.81	0.95	0.82	4.00E-04	0.0048	0.93	0.87	0.51	0.91	0.24	1.01	0.81	0.79	0.95	0.8
WG	-0.027	0.033	0.41	-	-	-	-0.00078	0.00094	0.4	-	-	-	0.021	0.15	0.89	-	-	-

Supplementary Table 4.3 Associations between heteroplasmic variants and obesity across sixteen mtDNA genes among AA

Gene	Definition 1						Definition 2						Definition 3					
	OR	95% CI	burden	skat	skato	acat	OR	95% CI	burden	skat	skato	acat	OR	95% CI	burden	skat	skato	acat
D-loop	0.89	(0.79, 0.99)	0.031	0.22	0.052	0.055	1	(0.99, 1)	0.027	0.05	0.07	0.035	0.24	(0.043, 1.32)	0.1	0.57	0.17	0.19
MT-RNR1	0.89	(0.67, 1.16)	0.38	0.8	0.55	0.64	1	(0.99, 1.01)	0.52	0.76	0.81	0.66	0.77	(0.44, 1.36)	0.37	0.75	0.55	0.59
MT-RNR2	0.86	(0.7, 1.05)	0.15	0.47	0.23	0.24	0.99	(0.99, 1)	0.1	0.3	0.33	0.16	0.69	(0.41, 1.14)	0.15	0.31	0.25	0.2
MT-ND1	0.89	(0.7, 1.12)	0.32	0.47	0.48	0.39	1	(0.99, 1)	0.26	0.67	0.46	0.44	0.74	(0.25, 2.21)	0.59	0.66	0.8	0.63
MT-ND2	1.11	(0.8, 1.53)	0.52	0.57	0.75	0.55	1	(0.99, 1.01)	0.45	0.52	0.78	0.49	0.83	(0.25, 2.79)	0.76	0.59	0.8	0.69
MT-CO1	0.85	(0.74, 0.97)	0.016	0.077	0.024	0.027	0.99	(0.99, 1)	0.01	0.18	0.025	0.02	0.87	(0.46, 1.62)	0.65	0.11	0.2	0.23
MT-CO2	0.57	(0.4, 0.81)	0.0015	0.49	0.0023	0.003	0.98	(0.97, 0.99)	0.0016	0.47	0.0077	0.0033	0.15	(0.039, 0.55)	0.0042	0.27	0.0075	0.0084
MT-ATP8	1.32	(0.69, 2.53)	0.4	0.16	0.27	0.23	1.01	(0.99, 1.03)	0.4	0.17	0.32	0.25	0.27	(0.0043, 16.24)	0.53	0.52	0.6	0.53
MT-ATP6	0.96	(0.72, 1.28)	0.8	0.21	0.36	0.51	1	(0.99, 1.01)	0.79	0.26	0.36	0.56	20.3	(0.17, 2372.41)	0.22	0.57	0.36	0.35
MT-CO3	0.72	(0.5, 1.04)	0.078	0.48	0.14	0.15	0.99	(0.98, 1)	0.082	0.38	0.19	0.14	0.6	(0.17, 2.1)	0.43	0.47	0.64	0.45
MT-ND3	1.24	(0.71, 2.17)	0.45	0.57	0.66	0.51	1.01	(0.99, 1.02)	0.36	0.63	0.77	0.49	0.98	(0.23, 4.25)	0.98	0.4	0.58	0.96
MT-ND4L	0.62	(0.33, 1.15)	0.13	0.22	0.2	0.16	0.99	(0.97, 1.01)	0.17	0.28	0.15	0.22	1.31	(0.21, 8.14)	0.77	0.25	0.38	0.52
MT-ND4	0.83	(0.65, 1.05)	0.11	0.58	0.18	0.22	0.99	(0.99, 1)	0.091	0.54	0.16	0.17	1.4	(0.6, 3.29)	0.44	0.31	0.5	0.37
MT-ND5	0.89	(0.77, 1.03)	0.11	0.3	0.16	0.17	1	(0.99, 1)	0.11	0.29	0.21	0.16	0.73	(0.36, 1.47)	0.38	0.22	0.36	0.29
MT-ND6	0.75	(0.52, 1.08)	0.13	0.16	0.21	0.14	0.99	(0.98, 1)	0.082	0.094	0.14	0.087	0.58	(0.1, 3.4)	0.55	0.095	0.16	0.18
MT-CYB	1.18	(0.93, 1.48)	0.17	0.41	0.28	0.25	1	(1, 1.01)	0.19	0.51	0.56	0.31	1.67	(0.47, 5.97)	0.43	0.79	0.65	0.66
WG	0.98	(0.96, 1)	0.058	-	-	-	1	(1, 1)	0.061	-	-	-	0.9	(0.78, 1.03)	0.12	-	-	-

Supplementary Table 4.4 Associations between heteroplasmic variants and obesity across sixteen mtDNA genes among EA

Gene	Definition 1							Definition 2							Definition 3						
	OR	95% CI	burden	skat	skato	acat		OR	95% CI	burden	skat	skato	acat		OR	95% CI	burden	skat	skato	acat	
D-loop	0.91	(0.8, 1.03)	0.14	1	0.25	1	1	(0.99, 1)	0.12	0.99	0.38	0.98	0.2	(0.037, 1.06)	0.058	0.96	0.11	0.83			
MT-RNR1	1.05	(0.81, 1.36)	0.71	0.77	0.9	0.75	1	(0.99, 1.01)	0.76	0.61	0.8	0.7	1.27	(0.78, 2.07)	0.33	0.47	0.53	0.39			
MT-RNR2	1.13	(0.91, 1.4)	0.26	0.83	0.43	0.61	1	(1, 1.01)	0.43	0.74	0.89	0.61	1.22	(0.83, 1.8)	0.32	0.79	0.51	0.6			
MT-ND1	1	(0.73, 1.37)	0.99	0.69	0.9	0.97	1	(0.99, 1.01)	0.97	0.81	0.8	0.95	0.8	(0.3, 2.1)	0.65	0.35	0.56	0.5			
MT-ND2	1.03	(0.76, 1.41)	0.84	0.39	0.61	0.71	1	(0.99, 1.01)	0.92	0.38	0.53	0.84	1.37	(0.52, 3.61)	0.52	0.4	0.62	0.46			
MT-CO1	1.14	(0.91, 1.42)	0.26	0.82	0.43	0.6	1	(1, 1.01)	0.3	0.65	0.23	0.46	1.22	(0.73, 2.04)	0.45	0.21	0.36	0.3			
MT-CO2	1.24	(0.91, 1.7)	0.18	0.14	0.25	0.16	1	(1, 1.01)	0.28	0.25	0.25	0.26	1.07	(0.44, 2.57)	0.88	0.093	0.17	0.39			
MT-ATP8	1.35	(0.73, 2.5)	0.34	0.17	0.29	0.23	1.01	(1, 1.02)	0.13	0.054	0.26	0.077	4.6	(0.01, 2123.04)	0.63	0.5	0.64	0.56			
MT-ATP6	0.96	(0.72, 1.29)	0.78	0.47	0.7	0.67	1	(0.99, 1.01)	0.96	0.41	0.26	0.92	2.44	(0.057, 104.94)	0.64	0.41	0.64	0.53			
MT-CO3	0.94	(0.69, 1.28)	0.68	0.84	0.89	0.78	1	(0.99, 1.01)	0.62	0.85	0.95	0.78	1.86	(0.66, 5.25)	0.24	0.41	0.4	0.31			
MT-ND3	1.03	(0.61, 1.73)	0.91	0.46	0.7	0.82	1	(0.99, 1.01)	0.93	0.29	0.11	0.83	0.62	(0.17, 2.25)	0.47	0.88	0.67	0.78			
MT-ND4L	1.25	(0.68, 2.29)	0.48	0.13	0.22	0.22	1	(0.99, 1.02)	0.56	0.18	0.056	0.3	1.88	(0.56, 6.29)	0.3	0.12	0.21	0.18			
MT-ND4	0.87	(0.69, 1.11)	0.27	0.98	0.43	0.96	1	(0.99, 1)	0.12	0.99	0.59	0.97	0.78	(0.38, 1.58)	0.49	0.35	0.56	0.42			
MT-ND5	0.92	(0.77, 1.09)	0.33	0.96	0.52	0.92	1	(0.99, 1)	0.19	0.84	0.054	0.56	0.75	(0.4, 1.39)	0.36	0.98	0.56	0.97			
MT-ND6	1.23	(0.86, 1.75)	0.26	0.28	0.44	0.27	1	(0.99, 1.01)	0.68	0.33	0.52	0.51	1.54	(0.41, 5.84)	0.52	0.26	0.43	0.37			
MT-CYB	1.03	(0.86, 1.25)	0.73	0.79	0.92	0.77	1	(1, 1.01)	0.81	0.58	0.54	0.73	0.88	(0.32, 2.43)	0.81	0.89	0.96	0.86			
WG	1	(0.97, 1.04)	0.92	-	-	-	1	(1, 1)	0.95	-	-	-	1.08	(0.92, 1.26)	0.33	-	-	-			

Supplementary Table 4.5 Associations between heteroplasmic variants and adjusted SBP across sixteen mtDNA genes among AA

Gene	Definition 1						Definition 2						Definition 3					
	beta	se	burden	skat	skato	acat	beta	se	burden	skat	skato	acat	beta	se	burden	skat	skato	acat
D-loop	-0.53	0.51	0.29	0.78	0.45	0.56	-0.018	0.018	0.31	0.89	0.49	0.76	-4.53	7.9	0.57	0.94	0.78	0.89
MT-RNR1	-2.67	1.27	0.035	0.34	0.056	0.066	-0.098	0.04	0.013	0.43	0.15	0.027	-4.89	2.63	0.063	0.15	0.1	0.089
MT-RNR2	-1.07	0.94	0.26	0.82	0.39	0.6	-0.029	0.028	0.29	0.89	0.34	0.75	-3.74	2.35	0.11	0.92	0.19	0.64
MT-ND1	-0.072	1.12	0.95	0.51	0.72	0.9	0.0067	0.031	0.83	0.37	0.48	0.68	7.46	5.1	0.14	0.19	0.25	0.16
MT-ND2	0.59	1.47	0.69	0.47	0.7	0.59	0.01	0.037	0.78	0.61	0.61	0.71	-7.99	5.6	0.15	0.83	0.26	0.47
MT-CO1	0.89	0.64	0.17	0.1	0.15	0.13	0.03	0.022	0.17	0.37	0.24	0.24	3.05	2.94	0.3	0.51	0.48	0.4
MT-CO2	-1.67	1.57	0.29	0.77	0.44	0.56	-0.037	0.045	0.42	0.91	0.73	0.83	3.01	6.21	0.63	0.79	0.83	0.73
MT-ATP8	-3.04	3.01	0.31	0.66	0.48	0.48	-0.08	0.078	0.3	0.79	0.84	0.6	-2.03	17.03	0.91	0.96	0.95	0.94
MT-ATP6	1.76	1.36	0.19	0.46	0.31	0.29	0.064	0.038	0.094	0.19	0.38	0.13	-8.05	22.9	0.73	0.75	0.91	0.74
MT-CO3	0.49	1.7	0.77	0.99	0.94	0.99	0.0072	0.045	0.87	0.97	0.96	0.96	-1.04	6.01	0.86	0.96	0.98	0.94
MT-ND3	-1.97	2.56	0.44	0.76	0.65	0.64	-0.06	0.074	0.41	0.71	0.93	0.58	-2.94	6.82	0.67	0.62	0.81	0.65
MT-ND4L	-5.47	2.89	0.059	0.15	0.091	0.085	-0.15	0.082	0.063	0.1	0.024	0.079	-10.39	8.41	0.22	0.67	0.32	0.4
MT-ND4	-0.3	1.1	0.79	0.8	0.94	0.8	-0.009	0.032	0.78	0.7	0.75	0.74	-4.41	3.79	0.24	0.32	0.4	0.28
MT-ND5	-0.64	0.66	0.33	0.86	0.47	0.71	-0.022	0.019	0.24	0.81	0.27	0.55	0.17	3.3	0.96	0.84	0.97	0.94
MT-ND6	0.14	1.72	0.94	0.79	0.94	0.9	-0.0074	0.05	0.88	0.85	0.75	0.87	-4.26	8.02	0.6	0.97	0.8	0.94
MT-CYB	-1.41	1.06	0.18	0.46	0.31	0.28	-0.031	0.028	0.28	0.7	0.047	0.48	-0.65	5.99	0.91	0.99	0.99	0.98
WG	-0.063	0.1	0.53	-	-	-	-0.0021	0.0031	0.51	-	-	-	-0.58	0.62	0.35	-	-	-

Supplementary Table 4.6 Associations between heteroplasmic variants and adjusted SBP across sixteen mtDNA genes among EA

Gene	Definition 1						Definition 2						Definition 3					
	beta	se	burden	skat	skato	acat	beta	se	burden	skat	skato	acat	beta	se	burden	skat	skato	acat
D-loop	-0.28	0.51	0.58	0.77	0.79	0.69	-0.0063	0.017	0.72	0.6	0.82	0.66	0.24	6.98	0.97	0.33	0.53	0.94
MT-RNR1	1.34	1.04	0.2	0.17	0.3	0.19	0.036	0.027	0.18	0.81	0.89	0.47	2.81	2.01	0.16	0.73	0.29	0.36
MT-RNR2	0.041	0.86	0.96	0.089	0.17	0.87	-0.0013	0.021	0.95	0.11	0.16	0.83	-0.88	1.58	0.58	0.41	0.64	0.49
MT-ND1	-0.37	1.25	0.77	0.45	0.68	0.64	-0.0045	0.028	0.87	0.73	0.52	0.82	-2.89	4.04	0.47	0.27	0.44	0.36
MT-ND2	0.072	1.25	0.95	0.33	0.54	0.9	-0.00056	0.029	0.98	0.34	0.15	0.97	3.49	3.67	0.34	0.44	0.53	0.39
MT-CO1	-1.26	0.89	0.16	0.89	0.28	0.64	-0.018	0.02	0.38	0.87	0.41	0.75	-1.87	2.14	0.38	0.87	0.6	0.75
MT-CO2	-0.83	1.24	0.5	0.79	0.73	0.68	-0.016	0.029	0.58	0.8	0.77	0.72	1.42	3.6	0.69	0.72	0.89	0.71
MT-ATP8	-2.11	2.38	0.37	0.85	0.57	0.71	-0.05	0.052	0.34	0.85	0.94	0.69	-3.99	17.21	0.82	0.95	0.91	0.91
MT-ATP6	0.43	1.16	0.71	1	0.91	0.99	0.0038	0.026	0.88	1	0.99	0.99	14.38	14.9	0.33	0.86	0.53	0.72
MT-CO3	0.14	1.25	0.91	0.02	0.037	0.05	0.0054	0.028	0.85	0.006	0.99	0.012	3.53	4.19	0.4	0.00051	0.00094	0.001
MT-ND3	-1.66	2.06	0.42	0.96	0.63	0.92	-0.03	0.049	0.54	0.78	0.96	0.69	-5.89	5.24	0.26	0.93	0.4	0.83
MT-ND4L	-0.23	2.54	0.93	0.86	0.97	0.9	0.0038	0.055	0.94	0.74	0.71	0.91	-3.98	5.16	0.44	0.84	0.65	0.72
MT-ND4	-0.53	0.95	0.58	0.69	0.79	0.64	-0.0071	0.024	0.77	0.83	0.83	0.8	1.7	2.93	0.56	0.46	0.69	0.51
MT-ND5	-0.57	0.7	0.41	0.62	0.62	0.52	-0.011	0.018	0.56	0.38	0.66	0.47	1.38	2.46	0.57	0.73	0.8	0.66
MT-ND6	0.93	1.38	0.5	0.74	0.73	0.64	0.014	0.032	0.65	0.89	0.96	0.83	2.61	4.95	0.6	0.4	0.61	0.5
MT-CYB	-0.38	0.74	0.6	0.69	0.82	0.65	0.0015	0.019	0.94	0.59	0.65	0.88	-2.07	3.96	0.6	0.24	0.41	0.39
WG	-0.089	0.13	0.5	-	-	-	-0.0017	0.0039	0.66	-	-	-	0.12	0.62	0.84	-	-	-

Supplementary Table 4.7 Associations between heteroplasmic variants and HTN across sixteen mtDNA genes among AA

Gene	Definition 1						Definition 2						Definition 3					
	OR	95% CI	burden	skat	skato	acat	OR	95% CI	burden	skat	skato	acat	OR	95% CI	burden	skat	skato	acat
D-loop	0.94	(0.82, 1.08)	0.37	0.5	0.56	0.44	1	(0.99, 1)	0.38	0.65	0.37	0.52	0.31	(0.04, 2.42)	0.26	0.37	0.42	0.31
MT-RNR1	0.71	(0.5, 1.01)	0.057	0.63	0.097	0.12	0.99	(0.98, 1)	0.056	0.73	0.48	0.13	0.51	(0.25, 1.03)	0.059	0.54	0.1	0.12
MT-RNR2	1.08	(0.84, 1.39)	0.56	0.79	0.77	0.7	1	(1, 1.01)	0.37	0.71	0.89	0.56	1.41	(0.77, 2.59)	0.27	0.61	0.44	0.42
MT-ND1	1.09	(0.81, 1.46)	0.56	0.73	0.77	0.66	1	(0.99, 1.01)	0.6	0.81	0.76	0.74	1.41	(0.41, 4.88)	0.59	0.56	0.78	0.58
MT-ND2	1.35	(0.93, 1.94)	0.11	0.6	0.2	0.22	1.01	(1, 1.02)	0.16	0.5	0.29	0.27	1.77	(0.48, 6.47)	0.39	0.52	0.59	0.45
MT-CO1	1.13	(0.96, 1.32)	0.15	0.043	0.064	0.067	1	(1, 1.01)	0.17	0.15	0.12	0.16	1.23	(0.59, 2.53)	0.58	0.34	0.54	0.46
MT-CO2	0.82	(0.54, 1.24)	0.34	0.6	0.51	0.46	0.99	(0.98, 1.01)	0.38	0.77	0.45	0.62	0.78	(0.17, 3.67)	0.75	0.53	0.75	0.66
MT-ATP8	0.93	(0.43, 2.01)	0.86	0.28	0.45	0.68	1	(0.98, 1.02)	0.87	0.33	0.8	0.73	0.14	(0.00021, 96.97)	0.56	0.7	0.67	0.63
MT-ATP6	1.25	(0.9, 1.74)	0.18	0.73	0.3	0.4	1.01	(1, 1.02)	0.048	0.48	0.21	0.094	0.85	(0.0054, 132.69)	0.95	0.16	0.28	0.86
MT-CO3	1.19	(0.79, 1.8)	0.4	0.57	0.61	0.48	1	(0.99, 1.02)	0.47	0.28	0.31	0.36	1.55	(0.35, 6.78)	0.56	0.8	0.78	0.71
MT-ND3	1.29	(0.67, 2.45)	0.45	0.82	0.65	0.69	1.01	(0.99, 1.03)	0.43	0.79	0.71	0.65	1.88	(0.34, 10.47)	0.47	0.41	0.6	0.44
MT-ND4L	0.74	(0.35, 1.54)	0.41	0.67	0.59	0.55	0.99	(0.97, 1.01)	0.46	0.49	0.25	0.48	0.75	(0.099, 5.62)	0.78	0.69	0.86	0.74
MT-ND4	0.9	(0.69, 1.18)	0.46	0.8	0.67	0.68	1	(0.99, 1)	0.42	0.55	0.81	0.48	0.39	(0.15, 1.03)	0.056	0.34	0.098	0.1
MT-ND5	1	(0.84, 1.18)	0.95	0.7	0.88	0.92	1	(0.99, 1)	0.95	0.52	0.57	0.91	1.06	(0.47, 2.4)	0.88	0.91	0.98	0.9
MT-ND6	0.88	(0.58, 1.32)	0.53	0.36	0.56	0.44	1	(0.98, 1.01)	0.45	0.64	0.6	0.55	0.31	(0.046, 2.05)	0.22	0.66	0.36	0.4
MT-CYB	0.77	(0.58, 1.01)	0.062	0.14	0.11	0.086	1	(0.99, 1)	0.22	0.23	0.02	0.22	0.75	(0.17, 3.31)	0.7	0.33	0.53	0.52
WG	1	(0.97, 1.03)	1	-	-	-	1	(1, 1)	0.89	-	-	-	1	(0.84, 1.18)	0.99	-	-	-

Supplementary Table 4.8 Associations between heteroplasmic variants and HTN across sixteen mtDNA genes among EA

Gene	Definition 1						Definition 2						Definition 3					
	OR	95% CI	burden	skat	skato	acat	OR	95% CI	burden	skat	skato	acat	OR	95% CI	burden	skat	skato	acat
D-loop	0.98	(0.86, 1.11)	0.71	0.28	0.45	0.49	1	(0.99, 1)	0.53	0.12	0.17	0.23	0.84	(0.18, 3.88)	0.83	0.5	0.73	0.72
MT-RNR1	1.19	(0.92, 1.52)	0.18	0.9	0.31	0.69	1	(1, 1.01)	0.11	0.89	0.47	0.53	1.49	(0.94, 2.36)	0.093	0.94	0.17	0.75
MT-RNR2	0.99	(0.8, 1.21)	0.89	0.21	0.36	0.69	1	(0.99, 1)	0.83	0.028	0.055	0.065	0.91	(0.63, 1.31)	0.61	0.15	0.27	0.29
MT-ND1	1.06	(0.78, 1.44)	0.7	0.11	0.2	0.24	1	(1, 1.01)	0.34	0.82	0.76	0.66	0.79	(0.31, 1.98)	0.61	0.18	0.31	0.33
MT-ND2	1.07	(0.79, 1.44)	0.67	0.53	0.76	0.61	1	(1, 1.01)	0.61	0.46	0.21	0.54	1.2	(0.45, 3.19)	0.72	0.78	0.9	0.75
MT-CO1	0.95	(0.76, 1.18)	0.63	0.71	0.85	0.67	1	(1, 1.01)	0.78	0.53	0.84	0.68	1.02	(0.62, 1.69)	0.94	0.68	0.89	0.89
MT-CO2	0.81	(0.59, 1.11)	0.19	0.44	0.33	0.28	0.99	(0.99, 1)	0.08	0.26	0.48	0.12	0.66	(0.27, 1.59)	0.36	0.43	0.56	0.39
MT-ATP8	0.87	(0.47, 1.64)	0.68	0.78	0.88	0.74	1	(0.98, 1.01)	0.72	0.67	0.71	0.69	0.24	(2.5e-05, 2339.8)	0.76	0.9	0.91	0.86
MT-ATP6	1.1	(0.82, 1.48)	0.52	0.84	0.76	0.74	1	(1, 1.01)	0.46	0.87	0.69	0.77	12.39	(0.26, 587.82)	0.2	0.87	0.34	0.63
MT-CO3	1.06	(0.78, 1.44)	0.7	0.93	0.9	0.88	1	(1, 1.01)	0.59	0.86	0.76	0.78	1.08	(0.41, 2.81)	0.88	0.21	0.36	0.67
MT-ND3	0.6	(0.36, 0.98)	0.042	0.11	0.075	0.061	0.99	(0.98, 1)	0.11	0.32	0.43	0.17	0.28	(0.08, 0.96)	0.042	0.091	0.072	0.058
MT-ND4L	1.33	(0.73, 2.42)	0.35	0.36	0.55	0.36	1.01	(0.99, 1.02)	0.27	0.5	0.77	0.37	1.58	(0.47, 5.31)	0.46	0.66	0.67	0.57
MT-ND4	1.02	(0.81, 1.28)	0.87	0.42	0.64	0.75	1	(1, 1.01)	0.49	0.69	0.78	0.6	1.38	(0.71, 2.71)	0.34	0.82	0.55	0.65
MT-ND5	0.95	(0.8, 1.13)	0.55	0.26	0.43	0.38	1	(1, 1)	0.74	0.27	0.42	0.51	1.17	(0.64, 2.13)	0.61	0.94	0.84	0.89
MT-ND6	1.19	(0.84, 1.69)	0.33	0.57	0.53	0.44	1	(1, 1.01)	0.34	0.83	0.8	0.67	1.42	(0.38, 5.29)	0.6	0.2	0.35	0.35
MT-CYB	0.86	(0.71, 1.03)	0.1	0.5	0.18	0.19	1	(0.99, 1)	0.23	0.75	0.24	0.47	0.54	(0.2, 1.46)	0.23	0.49	0.38	0.33
WG	0.99	(0.96, 1.03)	0.64	-	-	-	1	(1, 1)	0.99	-	-	-	1.01	(0.87, 1.17)	0.92	-	-	-

Supplementary Table 4.9 Associations between heteroplasmic variants and restricted BG across sixteen mtDNA genes among AA

Gene	Definition 1						Definition 2						Definition 3					
	beta	se	burden	skat	skato	acat	beta	se	burden	skat	skato	acat	beta	se	burden	skat	skato	acat
D-loop	0.51	0.26	0.051	0.068	0.084	0.058	0.014	0.0091	0.11	0.063	0.093	0.081	5.16	3.96	0.19	0.36	0.31	0.26
MT-RNR1	0.58	0.67	0.38	0.43	0.55	0.41	0.011	0.021	0.58	0.17	0.29	0.31	0.92	1.38	0.51	0.57	0.71	0.54
MT-RNR2	0.04	0.49	0.94	0.75	0.92	0.9	0.0052	0.014	0.72	0.65	0.84	0.69	-0.35	1.25	0.78	0.9	0.94	0.86
MT-ND1	-0.28	0.59	0.63	0.91	0.82	0.85	-0.01	0.016	0.52	0.91	0.71	0.84	-3.05	2.76	0.27	0.96	0.44	0.92
MT-ND2	0.62	0.78	0.43	0.025	0.045	0.048	0.011	0.019	0.56	0.081	0.14	0.16	2.23	2.91	0.44	0.37	0.56	0.41
MT-CO1	-0.094	0.32	0.77	0.32	0.46	0.57	-0.0063	0.011	0.57	0.52	0.73	0.55	-1.64	1.55	0.29	0.55	0.46	0.4
MT-CO2	-0.11	0.78	0.89	0.65	0.85	0.82	0.00034	0.022	0.99	0.62	0.81	0.98	1.52	2.99	0.61	0.094	0.16	0.19
MT-ATP8	0.069	1.55	0.96	0.88	0.98	0.94	0.0058	0.041	0.89	0.79	0.94	0.85	2.73	9.23	0.77	0.71	0.76	0.74
MT-ATP6	0.73	0.72	0.31	0.031	0.052	0.058	0.015	0.02	0.47	0.0049	0.0081	0.0098	-5.91	12.77	0.64	0.39	0.6	0.52
MT-CO3	-0.3	0.94	0.75	0.28	0.47	0.53	-0.017	0.025	0.48	0.34	0.53	0.41	-2.67	3.11	0.39	0.27	0.44	0.32
MT-ND3	1.29	1.48	0.39	0.1	0.17	0.17	0.02	0.041	0.63	0.042	0.069	0.088	1.38	3.62	0.7	0.12	0.19	0.27
MT-ND4L	0.51	1.48	0.73	0.64	0.83	0.69	0.013	0.043	0.76	0.32	0.49	0.57	-0.63	4.66	0.89	0.6	0.78	0.82
MT-ND4	0.18	0.61	0.76	0.067	0.11	0.16	0.012	0.017	0.5	0.02	0.039	0.039	1.74	2.05	0.4	0.053	0.099	0.098
MT-ND5	0.016	0.33	0.96	0.45	0.63	0.92	0.00016	0.0092	0.99	0.34	0.48	0.97	0.87	1.73	0.62	0.18	0.28	0.32
MT-ND6	0.95	0.94	0.31	0.42	0.47	0.36	0.032	0.026	0.22	0.64	0.28	0.38	1.01	4.34	0.82	0.46	0.67	0.7
MT-CYB	0.047	0.56	0.93	0.63	0.85	0.88	-0.0016	0.015	0.92	0.65	0.86	0.86	-0.78	3.15	0.8	0.61	0.83	0.73
WG	0.033	0.051	0.51	-	-	-	7.00E-04	0.0016	0.66	-	-	-	0.042	0.32	0.9	-	-	-

Supplementary Table 4.10 Associations between heteroplasmic variants and restricted BG across sixteen mtDNA genes among EA

Gene	Definition 1						Definition 2						Definition 3					
	beta	se	burden	skat	skato	acat	beta	se	burden	skat	skato	acat	beta	se	burden	skat	skato	acat
D-loop	-0.18	0.25	0.47	0.83	0.68	0.72	-0.0098	0.0083	0.24	0.53	0.42	0.36	-0.2	3.34	0.95	0.4	0.61	0.9
MT-RNR1	0.2	0.52	0.69	0.12	0.22	0.27	0.0081	0.013	0.54	0.27	0.43	0.38	0.95	0.99	0.34	0.045	0.084	0.082
MT-RNR2	-0.079	0.44	0.86	0.17	0.31	0.56	-0.0011	0.01	0.91	0.27	0.45	0.79	0.016	0.77	0.98	0.27	0.47	0.97
MT-ND1	-1.39	0.62	0.025	0.51	0.047	0.05	-0.033	0.014	0.016	0.41	0.028	0.032	-1.86	2.08	0.37	0.11	0.19	0.17
MT-ND2	-0.006	0.61	0.99	0.54	0.78	0.98	0.002	0.014	0.89	0.77	0.94	0.85	0.066	1.89	0.97	0.79	0.95	0.95
MT-CO1	0.77	0.44	0.078	0.18	0.14	0.11	0.015	0.0099	0.12	0.14	0.19	0.13	1	1.03	0.33	0.27	0.45	0.3
MT-CO2	-0.71	0.62	0.25	0.15	0.26	0.19	-0.016	0.014	0.25	0.15	0.27	0.19	0.38	1.77	0.83	0.47	0.69	0.72
MT-ATP8	-1.38	1.19	0.25	0.0014	0.0026	0.0029	-0.015	0.026	0.56	0.24	0.39	0.36	4.33	10.2	0.67	0.44	0.55	0.57
MT-ATP6	0.37	0.58	0.52	0.96	0.76	0.92	0.013	0.013	0.31	0.83	0.51	0.64	2.75	7.44	0.71	0.97	0.9	0.94
MT-CO3	0.091	0.63	0.88	0.35	0.56	0.76	-0.0013	0.014	0.92	0.22	0.36	0.8	2.47	2.06	0.23	0.29	0.39	0.26
MT-ND3	0.4	1.04	0.7	0.074	0.14	0.17	0.004	0.024	0.87	0.38	0.54	0.74	2.58	2.62	0.33	0.069	0.12	0.12
MT-ND4L	0.44	1.24	0.72	0.95	0.91	0.91	0.012	0.027	0.65	0.93	0.86	0.88	3.36	2.55	0.19	0.59	0.31	0.33
MT-ND4	-0.35	0.48	0.46	0.2	0.34	0.3	-0.016	0.012	0.19	0.23	0.29	0.21	0.31	1.37	0.82	0.33	0.53	0.64
MT-ND5	-0.14	0.37	0.7	0.31	0.51	0.51	-0.007	0.0089	0.43	0.54	0.56	0.49	-0.49	1.22	0.69	0.48	0.71	0.59
MT-ND6	-0.78	0.7	0.27	0.0054	0.01	0.011	-0.0091	0.016	0.56	0.054	0.098	0.11	-2.08	2.65	0.43	0.3	0.48	0.36
MT-CYB	0.46	0.37	0.22	0.029	0.054	0.052	0.013	0.0092	0.16	0.073	0.14	0.1	1.34	2	0.5	0.51	0.73	0.51
WG	-0.021	0.067	0.75	-	-	-	-0.00093	0.0019	0.62	-	-	-	0.27	0.31	0.39	-	-	-

Supplementary Table 4.11 Associations between heteroplasmic variants and diabetes across sixteen mtDNA genes among AA

Gene	Definition 1						Definition 2						Definition 3					
	OR	95% CI	burden	skat	skato	acat	OR	95% CI	burden	skat	skato	acat	OR	95% CI	burden	skat	skato	acat
D-loop	0.85	(0.72, 1.01)	0.07	0.72	0.12	0.16	0.99	(0.99, 1)	0.11	0.51	0.19	0.21	0.092	(0.0067, 1.26)	0.074	0.63	0.13	0.15
MT-RNR1	0.96	(0.65, 1.41)	0.84	0.7	0.89	0.78	1	(0.98, 1.01)	0.57	0.54	0.60	0.55	0.83	(0.39, 1.77)	0.63	0.47	0.69	0.55
MT-RNR2	0.87	(0.65, 1.17)	0.37	0.56	0.55	0.46	1	(0.99, 1.01)	0.67	0.34	0.69	0.50	0.83	(0.41, 1.67)	0.61	0.59	0.81	0.6
MT-ND1	0.96	(0.69, 1.35)	0.82	0.35	0.54	0.65	1	(0.99, 1.01)	0.6	0.26	0.40	0.41	0.74	(0.16, 3.44)	0.7	0.72	0.89	0.71
MT-ND2	1.15	(0.74, 1.79)	0.54	0.019	0.037	0.038	1	(0.99, 1.01)	0.77	0.27	0.62	0.54	0.9	(0.18, 4.54)	0.9	0.16	0.27	0.68
MT-CO1	0.86	(0.7, 1.06)	0.16	0.56	0.25	0.29	1	(0.99, 1)	0.26	0.38	0.21	0.31	1.01	(0.44, 2.34)	0.97	0.37	0.58	0.94
MT-CO2	0.66	(0.41, 1.07)	0.088	1	0.14	0.99	0.98	(0.97, 1)	0.019	0.98	0.16	0.69	0.13	(0.022, 0.81)	0.029	1	0.05	0.99
MT-ATP8	0.80	(0.34, 1.9)	0.62	0.49	0.7	0.55	1	(0.97, 1.02)	0.79	0.46	0.46	0.67	15.29	(0.00066, 354737.57)	0.59	0.59	0.75	0.59
MT-ATP6	1.14	(0.76, 1.71)	0.54	0.23	0.4	0.35	1	(0.99, 1.02)	0.52	0.32	0.55	0.41	56.42	(0.13, 24315.2)	0.19	0.39	0.32	0.27
MT-CO3	1.02	(0.64, 1.63)	0.93	0.55	0.78	0.87	1	(0.99, 1.01)	0.98	0.64	0.79	0.96	0.96	(0.21, 4.37)	0.95	0.42	0.64	0.91
MT-ND3	1.75	(0.86, 3.57)	0.12	0.22	0.2	0.16	1.01	(0.99, 1.03)	0.34	0.15	0.29	0.21	0.88	(0.12, 6.4)	0.9	0.78	0.93	0.86
MT-ND4L	0.64	(0.27, 1.54)	0.32	0.42	0.48	0.37	0.99	(0.96, 1.01)	0.28	0.7	0.12	0.48	0.57	(0.054, 5.94)	0.64	0.55	0.74	0.59
MT-ND4	1.07	(0.78, 1.48)	0.67	0.31	0.49	0.49	1	(0.99, 1.01)	0.79	0.44	0.50	0.66	0.82	(0.29, 2.33)	0.7	0.64	0.85	0.68
MT-ND5	0.95	(0.77, 1.18)	0.67	0.15	0.25	0.31	1	(0.99, 1)	0.63	0.26	0.44	0.42	1.66	(0.61, 4.49)	0.32	0.046	0.082	0.084
MT-ND6	1.03	(0.63, 1.69)	0.89	0.15	0.25	0.63	1	(0.98, 1.01)	0.92	0.3	0.17	0.82	0.48	(0.054, 4.2)	0.5	0.47	0.68	0.49
MT-CYB	0.86	(0.63, 1.17)	0.35	0.2	0.34	0.26	1	(0.99, 1.01)	0.55	0.17	0.55	0.30	0.73	(0.14, 3.84)	0.71	0.34	0.54	0.53
WG	0.99	(0.95, 1.02)	0.43	-	-	-	1	(1, 1)	0.45	-	-	-	0.93	(0.76, 1.15)	0.51	-	-	-

Supplementary Table 4.12 Associations between heteroplasmic variants and diabetes across sixteen mtDNA genes among EA

Gene	Definition 1				Definition 2				Definition 3								
	OR	95% CI	burden	skat	skato	acat	OR	95% CI	burden	skat	skato	acat	OR	95% CI	burden	skat	skato
D-loop	1.21(1.01, 1.46)	0.042	0.45	0.075	0.08	1	(1, 1.01)	0.17	0.26	0.44	0.21	9.58	(0.87, 105.32)	0.065	0.16	0.12	0.093
MT-RNR1	1.36(0.94, 1.97)	0.1	0.18	0.18	0.13	1.01	(1, 1.02)	0.072	0.0025	0.29	0.0048	1.8	(0.94, 3.46)	0.079	0.0055	0.011	0.01
MT-RNR2	0.99(0.75, 1.3)	0.93	0.52	0.75	0.87	1	(0.99, 1.01)	0.82	0.61	0.9	0.74	0.87	(0.53, 1.42)	0.57	0.76	0.8	0.69
MT-ND1	1.21(0.76, 1.94)	0.42	0.12	0.23	0.2	1.01	(0.99, 1.02)	0.33	0.044	0.67	0.079	4.08	(0.95, 17.46)	0.058	0.029	0.053	0.039
MT-ND2	1.16(0.74, 1.82)	0.52	0.37	0.59	0.44	1	(0.99, 1.01)	0.38	0.19	0.24	0.26	1.26	(0.32, 4.91)	0.74	0.36	0.56	0.57
MT-CO1	0.97(0.71, 1.32)	0.85	0.26	0.44	0.64	1	(0.99, 1.01)	0.90	0.37	0.58	0.79	0.6	(0.31, 1.16)	0.13	0.85	0.23	0.45
MT-CO2	0.81(0.52, 1.27)	0.36	0.84	0.57	0.7	1	(0.99, 1.01)	0.38	0.69	0.58	0.54	0.74	(0.21, 2.65)	0.64	0.77	0.85	0.72
MT-ATP8	1.13(0.48, 2.65)	0.78	0.11	0.2	0.29	1	(0.99, 1.02)	0.75	0.12	0.11	0.3	618.9	(0.00089, 431507131.83)	0.35	0.2	0.32	0.26
MT-ATP6	0.88(0.56, 1.36)	0.56	0.78	0.79	0.7	1	(0.99, 1.01)	0.52	0.91	0.91	0.84	0.035	(0.00011, 11.41)	0.26	0.86	0.42	0.67
MT-CO3	1.26(0.8, 1.98)	0.32	0.011	0.021	0.021	1	(0.99, 1.01)	0.62	0.16	0.091	0.31	1.19	(0.3, 4.76)	0.81	0.2	0.34	0.51
MT-ND3	1.32(0.64, 2.73)	0.45	0.19	0.34	0.28	1.01	(0.99, 1.02)	0.52	0.15	0.2	0.25	0.96	(0.18, 5.23)	0.97	0.37	0.56	0.93
MT-ND4L	1.50(0.62, 3.65)	0.37	0.1	0.18	0.17	1.01	(0.99, 1.03)	0.26	0.039	0.053	0.069	1.69	(0.31, 9.15)	0.54	0.28	0.45	0.4
MT-ND4	1.29(0.91, 1.82)	0.16	0.066	0.12	0.094	1	(0.99, 1.01)	0.43	0.35	0.89	0.39	0.54	(0.21, 1.34)	0.18	0.98	0.32	0.96
MT-ND5	1.16(0.91, 1.47)	0.24	0.17	0.29	0.2	1	(1, 1.01)	0.65	0.41	0.31	0.53	1.03	(0.45, 2.36)	0.94	0.58	0.81	0.89
MT-ND6	1.04(0.63, 1.71)	0.89	0.55	0.78	0.81	1	(0.99, 1.01)	0.84	0.42	0.62	0.71	2.77	(0.43, 17.77)	0.28	0.17	0.3	0.22
MT-CYB	0.90(0.69, 1.18)	0.46	0.87	0.68	0.76	1	(0.99, 1)	0.42	0.9	0.38	0.81	0.9	(0.22, 3.67)	0.88	0.5	0.73	0.79
WG	1.03(0.98, 1.08)	0.23	-	-	-	1	(1, 1)	0.34	-	-	-	1.02	(0.83, 1.25)	0.83	-	-	-

Supplementary Table 4.13 Associations between heteroplasmic variants and adjusted LDL across sixteen mtDNA genes among AA

Gene	Definition 1						Definition 2						Definition 3					
	beta	se	burden	skat	skato	acat	beta	se	burden	skat	skato	acat	beta	se	burden	skat	skato	acat
D-loop	1	1.04	0.34	0.066	0.11	0.11	0.031	0.038	0.41	0.058	0.047	0.11	14.69	16.57	0.38	0.42	0.55	0.4
MT-RNR1	1.72	2.59	0.51	0.53	0.7	0.52	0.049	0.083	0.55	0.27	0.18	0.39	5.4	5.37	0.31	0.26	0.41	0.28
MT-RNR2	1.44	1.97	0.47	0.12	0.19	0.2	0.013	0.059	0.82	0.23	0.24	0.56	0.031	4.8	0.99	0.68	0.88	0.99
MT-ND1	2.19	2.25	0.33	0.35	0.49	0.34	0.058	0.066	0.38	0.25	0.35	0.31	-6.11	10.47	0.56	0.98	0.78	0.97
MT-ND2	0.36	3.06	0.91	0.59	0.82	0.84	-0.009	0.08	0.91	0.46	0.91	0.83	-8.46	11.25	0.45	0.91	0.67	0.83
MT-CO1	1.36	1.31	0.3	0.14	0.21	0.2	0.042	0.047	0.37	0.024	0.012	0.046	0.87	6.08	0.89	0.11	0.2	0.48
MT-CO2	4.57	3.35	0.17	0.066	0.11	0.097	0.089	0.098	0.36	0.44	0.91	0.4	7.93	12.74	0.53	0.26	0.42	0.37
MT-ATP8	3.5	6.25	0.58	0.99	0.79	0.98	0.1	0.17	0.54	0.96	0.58	0.93	38.5	35.07	0.27	0.39	0.31	0.33
MT-ATP6	-1.26	2.69	0.64	0.94	0.84	0.9	-0.048	0.079	0.54	0.72	0.89	0.64	-24.15	44.29	0.59	1	0.8	1
MT-CO3	0.37	3.41	0.91	0.18	0.31	0.75	0.027	0.095	0.78	0.5	0.39	0.67	0.96	11.96	0.94	0.77	0.94	0.9
MT-ND3	5.42	5.28	0.3	0.089	0.16	0.14	0.2	0.16	0.2	0.12	0.95	0.15	-21.47	13.59	0.11	0.41	0.18	0.19
MT-ND4L	7.72	5.94	0.19	0.14	0.21	0.16	0.28	0.17	0.11	0.037	0.29	0.055	21.86	17.04	0.2	0.0045	0.0062	0.0089
MT-ND4	0.14	2.21	0.95	0.11	0.18	0.82	0.011	0.067	0.87	0.42	0.93	0.76	-2.69	7.99	0.74	0.48	0.7	0.63
MT-ND5	2.5	1.35	0.065	0.15	0.095	0.09	0.082	0.039	0.037	0.072	0.14	0.049	9.45	6.71	0.16	0.084	0.14	0.11
MT-ND6	1.2	3.39	0.72	0.18	0.3	0.39	0.027	0.1	0.79	0.34	0.73	0.62	-15.08	15.95	0.34	0.89	0.53	0.77
MT-CYB	-0.99	2.2	0.65	0.78	0.86	0.73	-0.034	0.06	0.57	0.53	0.79	0.55	-25.51	12.19	0.036	0.56	0.067	0.074
WG	0.22	0.21	0.28	-	-	-	0.0066	0.0066	0.32	-	-	-	0.41	1.29	0.75	-	-	-

Supplementary Table 4.14 Associations between heteroplasmic variants and adjusted LDL across sixteen mtDNA genes among EA

Gene	Definition 1						Definition 2						Definition 3					
	beta	se	burden	skat	skato	acat	beta	se	burden	skat	skato	acat	beta	se	burden	skat	skato	acat
D-loop	-0.67	0.95	0.48	0.32	0.51	0.39	0.0016	0.032	0.96	0.77	0.91	0.93	-11.39	12.62	0.37	0.42	0.57	0.39
MT-RNR1	0.083	1.95	0.97	0.021	0.037	0.1	-0.045	0.049	0.36	0.44	0.51	0.4	-1.82	3.65	0.62	0.32	0.52	0.46
MT-RNR2	-1.48	1.59	0.35	0.93	0.56	0.85	-0.024	0.039	0.54	0.89	0.29	0.81	-3.28	2.88	0.25	0.91	0.42	0.78
MT-ND1	-2.93	2.37	0.22	0.5	0.37	0.32	-0.064	0.052	0.22	0.45	0.28	0.31	-7.85	7.42	0.29	0.5	0.47	0.38
MT-ND2	-2.55	2.31	0.27	0.39	0.45	0.32	-0.028	0.052	0.59	0.19	0.47	0.33	-8.55	7.09	0.23	0.94	0.38	0.86
MT-CO1	-4.89	1.65	0.0031	0.14	0.0058	0.0061	-0.11	0.038	0.0032	0.12	0.01	0.0062	-13.4	3.9	6.00E-04	0.068	0.0011	0.0012
MT-CO2	2.09	2.37	0.38	0.4	0.59	0.39	0.062	0.056	0.27	0.25	0.67	0.26	7.88	6.79	0.25	0.02	0.039	0.038
MT-ATP8	-7.94	4.54	0.081	0.57	0.14	0.16	-0.18	0.098	0.074	0.38	0.13	0.13	-28.57	39.97	0.47	0.93	0.59	0.86
MT-ATP6	0.038	2.21	0.99	0.86	0.98	0.97	-0.0024	0.05	0.96	0.86	0.61	0.94	-26.21	29.03	0.37	0.67	0.62	0.52
MT-CO3	0.28	2.34	0.91	0.79	0.95	0.87	0.028	0.053	0.6	0.73	0.75	0.67	2.15	7.75	0.78	0.59	0.81	0.71
MT-ND3	-1.62	3.83	0.67	0.59	0.82	0.63	-0.012	0.09	0.9	0.71	0.69	0.84	0.22	9.45	0.98	0.99	1	0.99
MT-ND4L	1.29	4.7	0.78	0.24	0.4	0.53	-0.0009	0.1	0.99	0.34	0.79	0.98	-13.02	9.7	0.18	0.9	0.29	0.71
MT-ND4	-2.54	1.77	0.15	0.95	0.25	0.86	-0.047	0.045	0.3	0.97	0.84	0.93	0.88	5.29	0.87	0.58	0.81	0.79
MT-ND5	-1.19	1.28	0.35	0.77	0.55	0.6	-0.037	0.033	0.26	0.46	0.74	0.35	0.28	4.54	0.95	0.38	0.59	0.9
MT-ND6	0.83	2.66	0.76	0.81	0.93	0.79	0.0061	0.06	0.92	0.66	0.85	0.86	7.88	9.92	0.43	0.74	0.64	0.61
MT-CYB	-0.49	1.4	0.73	0.98	0.91	0.96	0.0063	0.036	0.86	0.94	0.97	0.92	-1.08	7.41	0.88	0.96	0.98	0.94
WG	-0.34	0.25	0.17	-	-	-	-0.0087	0.0072	0.23	-	-	-	-2.34	1.15	0.042	-	-	-

Supplementary Table 4.15 Associations between heteroplasmic variants and hyperlipidemia across sixteen mtDNA genes among AA

Gene	Definition 1							Definition 2							Definition 3						
	OR	95% CI	burden	skat	skato	acat	1	OR	95% CI	burden	skat	skato	acat	OR	95% CI	burden	skat	skato	acat		
D-loop	0.96	(0.86, 1.08)	0.49	0.66	0.7	0.58	1	(0.99, 1)	0.35	0.46	0.3	0.4	0.61	(0.1, 3.58)	0.58	0.74	0.79	0.67			
MT-RNR1	0.98	(0.73, 1.31)	0.89	0.93	0.98	0.91	1	(0.99, 1.01)	0.86	0.82	0.92	0.84	0.9	(0.5, 1.63)	0.73	0.89	0.91	0.84			
MT-RNR2	0.89	(0.72, 1.11)	0.3	0.57	0.46	0.42	1	(0.99, 1)	0.21	0.47	0.71	0.31	0.82	(0.49, 1.37)	0.45	0.27	0.45	0.35			
MT-ND1	0.98	(0.77, 1.25)	0.86	0.62	0.83	0.79	1	(0.99, 1.01)	0.93	0.62	0.89	0.88	1.02	(0.33, 3.15)	0.97	0.66	0.87	0.94			
MT-ND2	0.90	(0.65, 1.24)	0.52	0.1	0.19	0.19	1	(0.99, 1)	0.29	0.37	0.47	0.33	0.92	(0.29, 2.95)	0.89	0.37	0.57	0.77			
MT-CO1	0.95	(0.82, 1.09)	0.44	0.67	0.6	0.56	1	(0.99, 1)	0.42	0.68	0.85	0.56	1.08	(0.57, 2.05)	0.81	0.66	0.87	0.75			
MT-CO2	0.85	(0.6, 1.22)	0.39	0.81	0.58	0.66	1	(0.99, 1.01)	0.41	0.72	0.88	0.58	0.87	(0.22, 3.35)	0.84	0.74	0.92	0.8			
MT-ATP8	0.96	(0.5, 1.85)	0.9	0.5	0.72	0.82	1	(0.98, 1.02)	0.92	0.55	0.62	0.85	0.35	(0.0053, 22.91)	0.62	0.55	0.63	0.59			
MT-ATP6	0.92	(0.69, 1.23)	0.59	0.5	0.72	0.54	1	(0.99, 1.01)	0.46	0.3	0.36	0.37	34.92	(0.32, 3761.15)	0.14	0.31	0.23	0.19			
MT-CO3	0.76	(0.53, 1.09)	0.13	0.44	0.23	0.22	1	(0.99, 1.01)	0.42	0.61	0.39	0.52	0.62	(0.17, 2.23)	0.47	0.21	0.35	0.31			
MT-ND3	1.35	(0.77, 2.38)	0.29	0.68	0.46	0.48	1.01	(1, 1.03)	0.15	0.61	0.16	0.28	1.28	(0.29, 5.66)	0.75	0.42	0.61	0.61			
MT-ND4L	1.04	(0.55, 1.97)	0.91	0.92	0.99	0.92	1	(0.98, 1.02)	0.81	0.86	0.95	0.84	0.94	(0.15, 5.87)	0.94	0.79	0.93	0.91			
MT-ND4	0.91	(0.72, 1.15)	0.44	0.39	0.59	0.41	1	(0.99, 1)	0.43	0.46	0.42	0.44	0.88	(0.38, 2.06)	0.78	0.57	0.79	0.69			
MT-ND5	0.99	(0.86, 1.14)	0.88	0.44	0.63	0.78	1	(1, 1)	0.95	0.35	0.41	0.9	1.29	(0.63, 2.61)	0.49	0.43	0.64	0.46			
MT-ND6	0.88	(0.62, 1.26)	0.49	0.87	0.7	0.76	1	(0.99, 1.01)	0.56	0.82	0.62	0.73	0.55	(0.1, 2.94)	0.49	0.46	0.67	0.48			
MT-CYB	0.91	(0.72, 1.16)	0.46	0.42	0.64	0.44	1	(0.99, 1)	0.43	0.57	0.43	0.5	0.35	(0.094, 1.29)	0.11	0.32	0.2	0.17			
WG	0.99	(0.97, 1.01)	0.38	-	-	-	1	(1, 1)	0.35	-	-	-	0.95	(0.83, 1.1)	0.49	-	-	-			

Supplementary Table 4.16 Associations between heteroplasmic variants and hyperlipidemia across sixteen mtDNA genes among EA

Gene	Definition 1				Definition 2				Definition 3									
	OR	95% CI	burden	skat	skato	acat	OR	95% CI	burden	skat	skato	acat	OR	95% CI	burden	skat	skato	acat
D-loop	0.95	(0.85, 1.07)	0.42	0.75	0.63	0.61	1	(0.99, 1)	0.35	0.22	0.35	0.27	0.96	(0.22, 4.25)	0.96	0.74	0.92	0.93
MT-RNR1	0.86	(0.67, 1.09)	0.2	0.29	0.34	0.24	1	(0.99, 1)	0.15	0.52	0.48	0.26	0.71	(0.45, 1.12)	0.14	0.65	0.24	0.28
MT-RNR2	0.89	(0.73, 1.08)	0.25	0.53	0.41	0.36	1	(0.99, 1)	0.51	0.64	0.91	0.58	0.78	(0.55, 1.11)	0.16	0.38	0.28	0.24
MT-ND1	0.69	(0.52, 0.93)	0.013	0.029	0.025	0.018	0.99	(0.99, 1)	0.01	0.022	0.038	0.014	0.69	(0.28, 1.67)	0.41	0.30	0.49	0.35
MT-ND2	0.87	(0.65, 1.15)	0.33	0.39	0.53	0.36	1	(0.99, 1)	0.58	0.49	0.86	0.54	0.59	(0.24, 1.46)	0.25	0.24	0.41	0.25
MT-CO1	0.70	(0.57, 0.86)	0.0007	0.14	0.14	0.0014	0.99	(0.99, 1)	0.0013	0.19	0.086	0.0025	0.28	(0.17, 0.46)	3.40E-07	0.02	0.018	6.90E-07
MT-CO2	0.90	(0.66, 1.21)	0.47	0.58	0.7	0.52	1	(0.99, 1.01)	0.79	0.51	0.73	0.68	1.04	(0.45, 2.42)	0.93	0.42	0.64	0.86
MT-ATP8	0.54	(0.31, 0.95)	0.033	0.38	0.059	0.064	0.99	(0.98, 1)	0.053	0.35	0.52	0.095	0.017	(8.8e-05, 3.41)	0.13	0.53	0.18	0.24
MT-ATP6	0.86	(0.65, 1.13)	0.28	0.56	0.47	0.4	1	(0.99, 1)	0.6	0.47	0.28	0.53	0.018	(0.00051, 0.66)	0.029	0.22	0.053	0.051
MT-CO3	1.05	(0.79, 1.4)	0.75	0.78	0.93	0.77	1	(1, 1.01)	0.37	0.78	0.96	0.62	1.22	(0.47, 3.15)	0.68	0.34	0.55	0.52
MT-ND3	1.04	(0.65, 1.67)	0.88	0.66	0.87	0.81	1	(0.99, 1.01)	0.68	0.43	0.27	0.56	0.84	(0.25, 2.82)	0.77	0.77	0.93	0.77
MT-ND4L	0.79	(0.45, 1.39)	0.42	0.35	0.55	0.38	0.99	(0.98, 1.01)	0.39	0.33	0.55	0.36	0.63	(0.2, 2)	0.44	0.66	0.64	0.56
MT-ND4	0.89	(0.72, 1.11)	0.32	0.28	0.46	0.3	1	(0.99, 1)	0.44	0.51	0.56	0.48	1.1	(0.57, 2.1)	0.78	0.54	0.78	0.69
MT-ND5	0.99	(0.84, 1.17)	0.92	0.37	0.58	0.84	1	(1, 1)	0.91	0.35	0.87	0.8	0.61	(0.35, 1.08)	0.092	0.30	0.17	0.14
MT-ND6	1.11	(0.8, 1.55)	0.53	0.49	0.72	0.51	1	(0.99, 1.01)	0.7	0.33	0.34	0.52	3.04	(0.9, 10.29)	0.073	0.75	0.13	0.17
MT-CYB	0.94	(0.79, 1.11)	0.45	0.45	0.67	0.45	1	(1, 1)	0.97	0.55	0.7	0.95	0.98	(0.39, 2.46)	0.97	0.73	0.91	0.94
WG	0.97	(0.94, 1)	0.049	-	-	-	1	(1, 1)	0.075	-	-	-	0.79	(0.68, 0.9)	0.00082	-	-	-

REFERENCES

1. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics* 101, 5–22.
2. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* 47, D1005–D1012.
3. Wallace, D.C. (2005). A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. *Annual Review of Genetics* 39, 359–407.
4. Voet, D., Voet, J.G., and Pratt, C.W. (2006). *Fundamentals of biochemistry : life at the molecular level.*(Hoboken, N.J.: Wiley).
5. Gray, M.W. (2012). Mitochondrial evolution. *Cold Spring Harbor Perspectives in Biology* 4, a011403.
6. Wallace, D.C. (2013). A mitochondrial bioenergetic etiology of disease. *Journal of Clinical Investigation* 123, 1405–1412.
7. 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.
8. Stewart, J.B., and Chinnery, P.F. (2015). The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nature Reviews. Genetics* 16, 530–542.
9. Nieschlag, E., Habenicht, U.-F., and Nieschlag, S. (1992). *Spermatogenesis--fertilization--contraception : molecular, cellular, and endocrine events in male reproduction.*(Berlin; New York: Springer-Verlag).
10. Ankel-Simons, F., and Cummins, J.M. (1996). Misconceptions about mitochondria and mammalian fertilization: implications for theories on human evolution. *Proceedings of the National Academy of Sciences of the United States of America* 93, 13859–13863.
11. Eyre-Walker, A., and Awadalla, P. (2001). Does human mtDNA recombine? *Journal of Molecular Evolution* 53, 430–435.

12. Singh, G., Pachouri, U.C., Khaidem, D.C., Kundu, A., Chopra, C., and Singh, P. (2015). Mitochondrial DNA Damage and Diseases. *F1000Research* 4, 176.
13. 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.
14. Brown, M.D., Voljavec, A.S., Lott, M.T., Torroni, A., Yang, C.C., and Wallace, D.C. (1992). Mitochondrial DNA complex I and III mutations associated with Leber's hereditary optic neuropathy. *Genetics* 130, 163–173.
15. Howell, N., McCullough, D.A., Kubacka, I., Halvorson, S., and Mackey, D. (1992). The sequence of human mtDNA: the question of errors versus polymorphisms. *American Journal of Human Genetics* 50, 1333–1340.
16. Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., and Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genetics* 23, 147.
17. van Oven, M., and Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation* 30, E386–394.
18. Behar, D.M., van Oven, M., Rosset, S., Metspalu, M., Loogvali, E.L., Silva, N.M., Kivisild, T., Torroni, A., and Villems, R. (2012). A "Copernican" reassessment of the human mitochondrial DNA tree from its root. *American Journal of Human Genetics* 90, 675–684.
19. Bandelt, H.J., Kloss-Brandstatter, A., Richards, M.B., Yao, Y.G., and Logan, I. (2014). The case for the continuing use of the revised Cambridge Reference Sequence (rCRS) and the standardization of notation in human mitochondrial DNA studies. *Journal of Human Genetics* 59, 66–77.
20. Kogelnik, A.M., Lott, M.T., Brown, M.D., Navathe, S.B., and Wallace, D.C. (1996). MITOMAP: a human mitochondrial genome database. *Nucleic Acids Research* 24, 177–179.
21. Wallace, D.C. (2015). Mitochondrial DNA variation in human radiation and disease. *Cell* 163, 33–38.
22. Wallace, D.C., and Chalkia, D. (2013). Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. *Cold Spring Harbor Perspectives in Biology* 5, a021220.

23. Chinnery, P.F., and Gomez-Duran, A. (2018). Oldies but Goldies mtDNA Population Variants and Neurodegenerative Diseases. *Frontiers in Neuroscience* 12, 682.
24. Ding, J., Sidore, C., Butler, T.J., Wing, M.K., Qian, Y., Meirelles, O., Busonero, F., Tsoi, L.C., 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 Genetics* 11, e1005306.
25. Liu, C., Fetterman, J.L., Liu, P., Luo, Y., Larson, M.G., Vasan, R.S., Zhu, J., and Levy, D. (2018). Deep sequencing of the mitochondrial genome reveals common heteroplasmic sites in NADH dehydrogenase genes. *Human Genetics* 137, 203–213.
26. Ye, K., Lu, J., Ma, F., Keinan, A., and Gu, Z. (2014). Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals. *Proceedings of the National Academy of Sciences of the United States of America* 111, 10654–10659.
27. Taylor, R.W., and Turnbull, D.M. (2005). Mitochondrial DNA mutations in human disease. *Nature Reviews. Genetics* 6, 389–402.
28. Saxena, R., de Bakker, P.I., Singer, K., Mootha, V., Burt, N., Hirschhorn, J.N., Gaudet, D., Isomaa, B., Daly, M.J., Groop, L., et al. (2006). Comprehensive association testing of common mitochondrial DNA variation in metabolic disease. *American Journal of Human Genetics* 79, 54–61.
29. Fetterman, J.L., Liu, C., Mitchell, G.F., Vasan, R.S., Benjamin, E.J., Vita, J.A., Hamburg, N.M., and Levy, D. (2018). Relations of mitochondrial genetic variants to measures of vascular function. *Mitochondrion* 40, 51–57.
30. Kraja, A.T., Liu, C., Fetterman, J.L., Graff, M., Have, C.T., Gu, C., Yanek, L.R., Feitosa, M.F., Arking, D.E., Chasman, D.I., et al. (2019). Associations of Mitochondrial and Nuclear Mitochondrial Variants and Genes with Seven Metabolic Traits. *American Journal of Human Genetics* 104, 112–138.
31. Liu, C., Dupuis, J., Larson, M.G., and Levy, D. (2013). Association testing of the mitochondrial genome using pedigree data. *Genetic Epidemiology* 37, 239–247.
32. Liu, C., Yang, Q., Hwang, S.J., Sun, F., Johnson, A.D., Shrihail, O.S., Vasan, R.S., Levy, D., and Schwartz, F. (2012). Association of genetic variation in

the mitochondrial genome with blood pressure and metabolic traits. *Hypertension* 60, 949–956.

33. van der Walt, J.M., Dementieva, Y.A., Martin, E.R., Scott, W.K., Nicodemus, K.K., Kroner, C.C., Welsh-Bohmer, K.A., Saunders, A.M., Roses, A.D., Small, G.W., et al. (2004). Analysis of European mitochondrial haplogroups with Alzheimer disease risk. *Neuroscience Letters* 365, 28–32.
34. Procaccio, V., Neckelmann, N., Paquis-Flucklinger, V., Bannwarth, S., Jimenez, R., Davila, A., Poole, J.C., and Wallace, D.C. (2006). Detection of low levels of the mitochondrial tRNA^{Leu}(UUR) 3243A>G mutation in blood derived from patients with diabetes. *Molecular Diagnosis & Therapy* 10, 381–389.
35. Zaragoza, M.V., Fass, J., Diegoli, M., Lin, D., and Arbustini, E. (2010). Mitochondrial DNA variant discovery and evaluation in human Cardiomyopathies through next-generation sequencing. *PLoS One* 5, e12295.
36. Sosa, M.X., Sivakumar, I.K., Maragh, S., Veeramachaneni, V., Hariharan, R., Parulekar, M., Fredrikson, K.M., Harkins, T.T., Lin, J., Feldman, A.B., et al. (2012). Next-generation sequencing of human mitochondrial reference genomes uncovers high heteroplasmy frequency. *PLoS Computational Biology* 8, e1002737.
37. Huang, T. (2011). Next-generation sequencing to characterize mitochondrial genomic DNA heteroplasmy. *Current Protocols in Human Genetics* Chapter 19, Unit19 18.
38. Zhang, W., Cui, H., and Wong, L.J. (2012). Comprehensive one-step molecular analyses of mitochondrial genome by massively parallel sequencing. *Clinical Chemistry* 58, 1322–1331.
39. Schmitt, M.W., Kennedy, S.R., Salk, J.J., Fox, E.J., Hiatt, J.B., and Loeb, L.A. (2012). Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences of the United States of America* 109, 14508–14513.
40. Ahn, E.H., Hirohata, K., Kohn, B.F., Fox, E.J., Chang, C.C., and Loeb, L.A. (2015). Detection of Ultra-Rare Mitochondrial Mutations in Breast Stem Cells by Duplex Sequencing. *PLoS One* 10, e0136216.
41. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021).

Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 590, 290–299.

42. Calabrese, C., Simone, D., Diroma, M.A., Santorsola, M., Gutta, C., Gasparre, G., Picardi, E., Pesole, G., and Attimonelli, M. (2014). MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. *Bioinformatics* 30, 3115–3117.
43. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 43, 491–498.
44. 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. *Current Protocols in Bioinformatics* 44(123), 1.23.1–26. <https://doi.org/10.1002/0471250953.bi0123s44>
45. Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R., and Lander, E.S. (2014). Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America* 111, E455–464.
46. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American Journal of Human Genetics* 83, 311–321.
47. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics* 89, 82–93.
48. Lee, S., Wu, M.C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 13, 762–775.
49. Liu, Y., Chen, S., Li, Z., Morrison, A.C., Boerwinkle, E., and Lin, X. (2019). ACAT: A Fast and Powerful p-value Combination Method for Rare-Variant Analysis in Sequencing Studies. *American Journal of Human Genetics* 104, 410–421.
50. Han, F., and Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. *Human Heredity* 70, 42–54.

51. Sha, Q., Wang, S., and Zhang, S. (2013). Adaptive clustering and adaptive weighting methods to detect disease associated rare variants. *European Journal of Human Genetics* 21, 332–337.
52. Liu, C., Fetterman, J.L., Qian, Y., Sun, X., Blackwell, T.W., Pitsillides, A., Cade, B.E., Wang, H., Raffield, L.M., Lange, L.A., et al. (2021). Presence and transmission of mitochondrial heteroplasmic mutations in human populations of European and African ancestry. *Mitochondrion* 60, 33–42.
53. Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *American Journal of Human Genetics* 95, 5–23.
54. (1989). The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *American Journal of Epidemiology* 129, 687–702.
55. Dawber, T.R., Meadors, G.F., and Moore, F.E., Jr. (1951). Epidemiological approaches to heart disease: the Framingham Study. *American Journal of Public Health Nations Health* 41, 279–281.
56. Feinleib, M., Kannel, W.B., Garrison, R.J., McNamara, P.M., and Castelli, W.P. (1975). The Framingham Offspring Study. Design and preliminary data. *Preventive Medicine* 4, 518–525.
57. Splansky, G.L., Corey, D., Yang, Q., Atwood, L.D., Cupples, L.A., Benjamin, E.J., D'Agostino, R.B., Sr., Fox, C.S., Larson, M.G., Murabito, J.M., et al. (2007). The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *American Journal of Epidemiology* 165, 1328–1335.
58. Fried, L.P., Borhani, N.O., Enright, P., Furberg, C.D., Gardin, J.M., Kronmal, R.A., Kuller, L.H., Manolio, T.A., Mittelmark, M.B., Newman, A., et al. (1991). The Cardiovascular Health Study: design and rationale. *Annals of Epidemiology* 1, 263–276.
59. Wilson, J.G., Rotimi, C.N., Ekunwe, L., Royal, C.D., Crump, M.E., Wyatt, S.B., Steffes, M.W., Adeyemo, A., Zhou, J., Taylor, H.A., Jr., et al. (2005). Study design for genetic analysis in the Jackson Heart Study. *Ethnicity & Disease* 15, S6-30-37.
60. Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R., Diez Roux, A.V., Folsom, A.R., Greenland, P., Jacob, D.R., Jr., Kronmal, R., Liu, K., et al. (2002). Multi-Ethnic Study of Atherosclerosis: objectives and design. *American Journal of Epidemiology* 156, 871–881.

61. Chen, H., Huffman, J.E., Brody, J.A., Wang, C., Lee, S., Li, Z., Gogarten, S.M., Sofer, T., Bielak, L.F., Bis, J.C., et al. (2019). Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies. *American Journal of Human Genetics* 104, 260–274.
62. Liu, X., Longchamps, R.J., Wiggins, K.L., Raffield, L.M., Bielak, L.F., Zhao, W., Pitsillides, A., Blackwell, T.W., Yao, J., Guo, X., et al. (2021). Association of mitochondrial DNA copy number with cardiometabolic diseases. *Cell Genomics* 1(1) 100006.
<https://doi.org/10.1016/j.xgen.2021.100006>
63. Fisher, R.A. (1925). *Statistical methods for research workers.*(Edinburgh, London,: Oliver and Boyd).
64. Borenstein, M., Hedges, L.V., Higgins, J.P., and Rothstein, H.R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods* 1, 97–111.
65. R Core Team. (2019). *R: A language and environment for statistical computing.* In. (R Foundation for Statistical Computing, Vienna, Austria).
66. Sondheimer, N., Glatz, C.E., Tirone, J.E., Deardorff, M.A., Krieger, A.M., and Hakonarson, H. (2011). Neutral mitochondrial heteroplasmy and the influence of aging. *Human Molecular Genetics* 20, 1653–1659.
67. Scheffler, I.E. (2008). *Mitochondria.*(Hoboken, N.J.: Wiley-Liss).
68. Li, R., Greinwald, J.H., Jr., Yang, L., Choo, D.I., Wenstrup, R.J., and Guan, M.X. (2004). Molecular analysis of the mitochondrial 12S rRNA and tRNASer(UCN) genes in paediatric subjects with non-syndromic hearing loss. *Journal of Medical Genetics* 41, 615–620.
69. Yao, Y.G., Salas, A., Bravi, C.M., and Bandelt, H.J. (2006). A reappraisal of complete mtDNA variation in East Asian families with hearing impairment. *Human Genetics* 119, 505–515.
70. Yen, K., Mehta, H.H., Kim, S.J., Lue, Y., Hoang, J., Guerrero, N., Port, J., Bi, Q., Navarrete, G., Brandhorst, S., et al. (2020). The mitochondrial derived peptide humanin is a regulator of lifespan and healthspan. *Aging (Albany NY)* 12, 11185–11199.
71. Tajima, H., Niikura, T., Hashimoto, Y., Ito, Y., Kita, Y., Terashita, K., Yamazaki, K., Koto, A., Aiso, S., and Nishimoto, I. (2002). Evidence for in vivo production of Humanin peptide, a neuroprotective factor against Alzheimer's disease-related insults. *Neuroscience Letters* 324, 227–231.

72. Fontanesi, F., Soto, I.C., and Barrientos, A. (2008). Cytochrome c oxidase biogenesis: new levels of regulation. *IUBMB Life* 60, 557–568.
73. Brown, M.D., Yang, C.C., Trounce, I., Torroni, A., Lott, M.T., and Wallace, D.C. (1992). A mitochondrial DNA variant, identified in Leber hereditary optic neuropathy patients, which extends the amino acid sequence of cytochrome c oxidase subunit I. *American Journal of Human Genetics* 51, 378–385.
74. Varlamov, D.A., Kudin, A.P., Vielhaber, S., Schroder, R., Sassen, R., Becker, A., Kunz, D., Haug, K., Rebstock, J., Heils, A., et al. (2002). Metabolic consequences of a novel missense mutation of the mtDNA CO I gene. *Human Molecular Genetics* 11, 1797–1805.
75. Capaldi, R.A. (1990). Structure and function of cytochrome c oxidase. *Annual Review of Biochemistry* 59, 569–596.
76. Lagerstrom-Fermer, M., Olsson, C., Forsgren, L., and Syvanen, A.C. (2001). Heteroplasmy of the human mtDNA control region remains constant during life. *American Journal of Human Genetics* 68, 1299–1301.
77. Sharma, L.K., Lu, J., and Bai, Y. (2009). Mitochondrial respiratory complex I: structure, function and implication in human diseases. *Current Medicinal Chemistry* 16, 1266–1277.
78. Junge, W., and Nelson, N. (2015). ATP synthase. *Annual Review of Biochemistry* 84, 631–657.
79. Saneto, R.P. (2020). Mitochondrial diseases: expanding the diagnosis in the era of genetic testing. *Journal of Translational Genetics and Genomics* 4, 384–428. <https://doi.org/10.20517/jtgg.2020.40>
80. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research* 45, D896–D901.
81. Surendran, P., Feofanova, E.V., Lahrouchi, N., Ntalla, I., Karthikeyan, S., Cook, J., Chen, L., Mifsud, B., Yao, C., Kraja, A.T., et al. (2020). Discovery of rare variants associated with blood pressure regulation through meta-analysis of 1.3 million individuals. *Nature Genetics* 52, 1314–1332.
82. Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society Series B (Methodological)* 39, 1–38.

83. Turkmen, A., and Lin, S. (2017). Are rare variants really independent? *Genetic Epidemiology* 41, 363–371.
84. Genomes Project, C., 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.
85. Meyer, H.V., and Birney, E. (2018). PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. *Bioinformatics* 34, 2951–2956.
86. Hubert, L., and Arabie, P. (1985). Comparing partitions. *Journal of Classification* 2, 193–218.
87. Ehret, G.B., and Caulfield, M.J. (2013). Genes for blood pressure: an opportunity to understand hypertension. *European Heart Journal* 34, 951–961.
88. Forouzanfar, M.H., Liu, P., Roth, G.A., Ng, M., Biryukov, S., Marczak, L., Alexander, L., Estep, K., Hassen Abate, K., Akinyemiju, T.F., et al. (2017). Global Burden of Hypertension and Systolic Blood Pressure of at Least 110 to 115 mm Hg, 1990–2015. *JAMA: The Journal of the American Medical Association* 317, 165–182.
89. Surendran, P., Feofanova, E.V., Lahrouchi, N., Ntalla, I., Karthikeyan, S., Cook, J., Chen, L., Mifsud, B., Yao, C., Kraja, A.T., et al. (2021). Publisher Correction: Discovery of rare variants associated with blood pressure regulation through meta-analysis of 1.3 million individuals. *Nature Genetics* 53, 762.
90. Liu, C., Kraja, A.T., Smith, J.A., Brody, J.A., Franceschini, N., Bis, J.C., Rice, K., Morrison, A.C., Lu, Y., Weiss, S., et al. (2016). Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci. *Nature Genetics* 48, 1162–1170.
91. Bomba, L., Walter, K., and Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome Biology* 18, 77.
92. Wallace, D.C., Fan, W., and Procaccio, V. (2010). Mitochondrial energetics and therapeutics. *Annual Review of Pathology* 5, 297–348.
93. Lee, Y.T., Lin, H.Y., Chan, Y.W., Li, K.H., To, O.T., Yan, B.P., Liu, T., Li, G., Wong, W.T., Keung, W., et al. (2017). Mouse models of atherosclerosis: a historical perspective and recent advances. *Lipids in Health and Disease* 16, 12.

94. Schaefer, A.M., McFarland, R., Blakely, E.L., He, L., Whittaker, R.G., Taylor, R.W., Chinnery, P.F., and Turnbull, D.M. (2008). Prevalence of mitochondrial DNA disease in adults. *Annals of Neurology* 63, 35–39.
95. Schwartz, F., Duka, A., Sun, F., Cui, J., Manolis, A., and Gavras, H. (2004). Mitochondrial genome mutations in hypertensive individuals. *American Journal of Hypertension* 17, 629–635.
96. Wei, W., Tuna, S., Keogh, M.J., Smith, K.R., Aitman, T.J., Beales, P.L., Bennett, D.L., Gale, D.P., Bitner-Glindzicz, M.A.K., Black, G.C., et al. (2019). Germline selection shapes human mitochondrial DNA diversity. *Science* 364(6442), eaau6520. <https://doi.org/10.1126/science.aau6520>
97. Friedman, G.D., Cutter, G.R., Donahue, R.P., Hughes, G.H., Hulley, S.B., Jacobs, D.R., Jr., Liu, K., and Savage, P.J. (1988). CARDIA: study design, recruitment, and some characteristics of the examined subjects. *Journal of Clinical Epidemiology* 41, 1105–1116.
98. Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R., Diez Roux, A.V., Folsom, A.R., Greenland, P., Jacobs Jr., D.R., Kronmal, R., Liu, K., et al. (2002). Multi-Ethnic Study of Atherosclerosis: Objectives and Design. *American Journal of Epidemiology* 156, 871–881.
99. (2021). TOPMed Whole Genome Sequencing Methods: Freeze 8. <https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-methods-freeze-8>
100. Battle, S.L., Puiu, D., Group, T.O.m.W., Verlow, J., Broer, L., Boerwinkle, E., Taylor, K.D., Rotter, J.I., Rich, S.S., Grove, M.L., et al. (2022). A bioinformatics pipeline for estimating mitochondrial DNA copy number and heteroplasmy levels from whole genome sequencing data. *NAR Genomics and Bioinformatics* 4, lqac034.
101. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al. (2013). Discovery and refinement of loci associated with lipid levels. *Nature Genetics* 45, 1274–1283.
102. Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R., and Yoshikawa, S. (1996). The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 Å. *Science* 272, 1136–1144.
103. Zou, Y., Carbonetto, P., Wang, G., and Stephens, M. (2022). Fine-mapping from summary data with the "Sum of Single Effects" model. *PLoS Genetics* 18, e1010299.

CURRICULUM VITAE

